nature

23 November 1978

Prejudicial to the interests of the state?

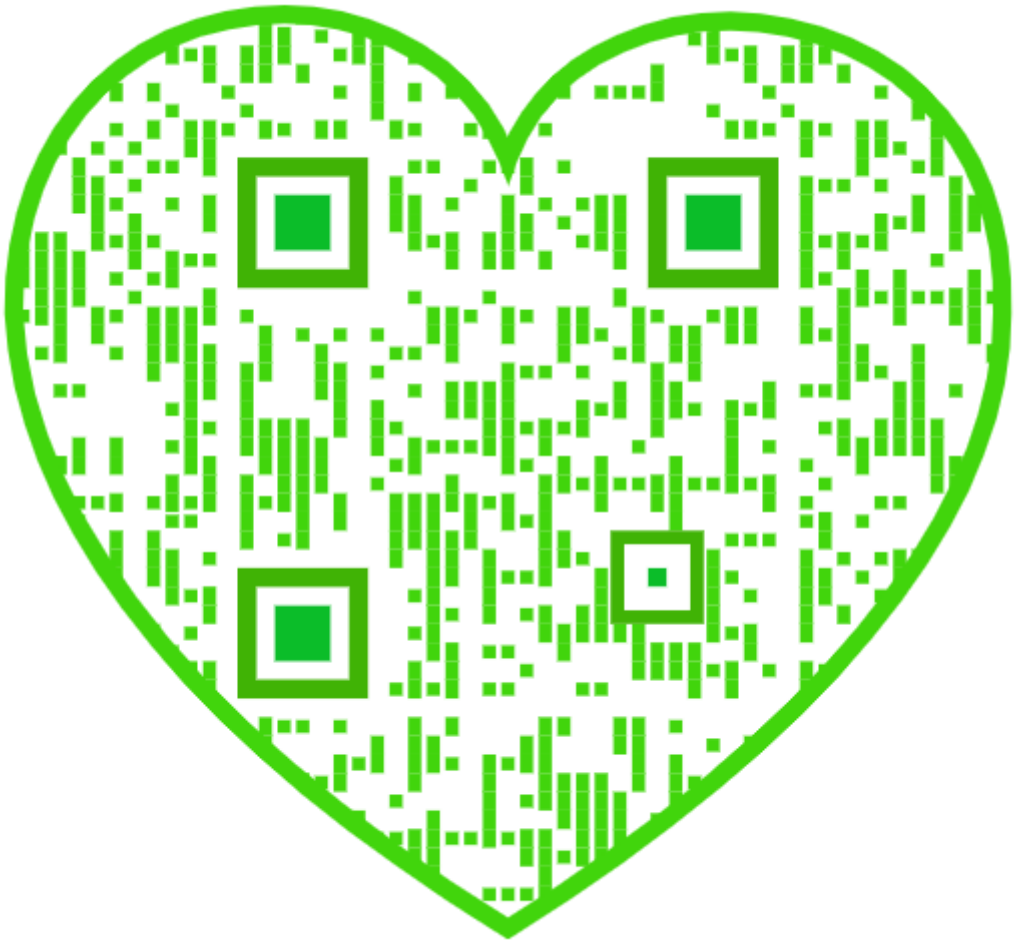
FOR nearly eighteen months the British courts have been occupied in one way or another with the cases of Crispin Aubrey, Duncan Campbell and John Berry, who had been accused of various breaches of the Official Secrets Act. Mr Berry is a former analyst in an overseas signals intelligence unit, and Messrs Aubrey and Campbell are journalists who recorded a three-hour conversation with Mr Berry on the subject of signals intelligence in February 1977. Immediately after the conversation all three were arrested.

A variety of charges were directed at the defendants, most notably under section 1 of the 1911 Official Secrets Act which prescribes up to 14 years in prison for those convicted of passing information for a "purpose prejudicial to the safety and interests of the state". Past prosecutions under this section have been confined to cases of espionage and sabotage; the purposes of the defendants seem to have been limited to the dissemination of information through journalism so a conviction would certainly have broken new ground in interpretation of the act. In the event, as proceedings wore on, with many peculiar asides down to the revelation that the foreman of one jury was himself a former soldier with the highly secretive Special Air Services Regiment, the Section 1 charges were dropped and the defendants were only accused under Section 2 of the act, which makes it an offence simply to transmit or receive any unauthorised information about what government employees do, see or make. Success for the prosecution under those circumstances was inevitable; there cannot be a single self-respecting journalist, nor for that matter government employee, who could not be prosecuted under Section 2 for some indiscretion or other. The judge, having earlier hinted that he did not intend to impose immediate prison sentences, lived up to his promise, sentencing Mr Berry only to a suspended six months in gaol and providing Messrs Aubrey and Campbell with conditional discharges.

It is important that this case be remembered for more than its elements of farce and anticlimax, and in particular a perspective has to be recognised on the wider issues involved. True, Mr Berry did pass on rather low-grade information concerning secret activities of the British government in the accumulation of intelligence—information which, it is claimed, is in any case almost in the public domain. And true there must be some small aspect of a government's activities which, for the sake of national security, is best not widely known (though Mr Berry's revelations do not seem, in any sense, to fall into such a category). But is it worth a quarter of a million pounds of public money to try and crush perpetrations of peccadilloes which could have been dealt with summarily in a magistrates' court? Maybe the judge's lenient sentences will have put heart into those who wish to reform our whole attitude to secrecy in government—a long-delayed activity. More likely the heavyweight support the government has shown for the maintenance of a *cordon sanitaire* around its activities will encourage those in government service who hold the view that nobody outside a charmed circle should know anything at all.

The Official Secrets Act casts a very long shadow. The same act that says it is undesirable that the current whereabouts of Britain's nuclear submarines should be widely known also says that any difference of opinion within a government agency is not for outside consumption. And the penumbra is even larger; those within the confines of the Officials Secrets Act are expected to develop a profound group loyalty which regards even such a harmless thing as a joking reference in print to the organisation as meriting official displeasure.

The results of decades of obsessive British secrecy and the harbouring of inside knowledge are around for all to see. In the Western world there cannot be a single government less open in its intentions with the public, and more distrustful of those, like journalists, whose profession is not just to repeat the official line but to try and shed light on the inner processes of decision-making. And the lack of easy flow of qualified people into and out of government service just reinforces the insider mentality. It is a total change of attitude, not a rewriting of the Official Secrets Act, that is needed to open up government—and this may take a generation. But by its vigorous pursuit of minor journalistic escapades the British government has shown it doesn't yet recognise the need for change at all. □



Inflation–and Congress–kill growth in funds for basic science

President Carter has had difficulty persuading Congress to increase support for basic research. Next year promises to be even harder. **David Dickson** reports.

Now that the dust is beginning to settle, it appears that apart from biomedical research President Carter has failed to deliver one of the promises made in his budget request to Congress for the financial year 1979, namely a "real growth" in funds for basic science.

The President had asked for a total of $3.6 billion, which would have resulted in an overall funding level 5 per cent above inflation. But the request has fallen victim to two factors: a higher-than-expected inflation rate, and a Congress reacting strongly to grass-roots demands for general reductions in federal spending.

Both factors continue to dominate planning for the 1980 budget, which will be announced in January. And despite indications from the president that he wants to see basic research protected from the most severe cuts, few expect that it will receive any particular privileges, or escape from the general constraints being imposed.

Speaking at a news conference earlier this month, for example, the President merely said that: "I have directed in the preparation of the 1980 fiscal year budget that basic research and research and development in general *should not be reduced as a percentage of the total federal budget*" (emphasis added).

The precise effects of Congressional actions on the science budget over the past year have yet to be officially calculated. The one bright spot—which tends to make the overall picture appear better than it really is—is a 34 per cent increase in the funds available for basic biomedical research through the National Institutes of Health.

Apart from this, officials of the American Association for the Advancement of Science have calculated the funds for basic science in all other fields will be 9.3 per cent higher than 1978. Dr Richard Atkinson, director of the National Science Foundation, gave a figure last week of 8.3 per cent; and others expect the figure to be even lower—close to the inflation rate.

The result has been a major disappointment to the scientific community. "The President's request for real growth in the science budget was met with great delight by the academic community; we are obviously disappointed by the lack of significant real growth in the budget", Dr Gerald Lieberman, vice provost and dean of research at Stanford said last week.

Overall, funds for research and development have not done too badly. The President asked Congress for an increase in R&D expenditure of 6.8 per cent, to a total of $29.3 billion. This figure was increased by Congress by a further $315.5 million, the major growth occurring in the Department of Energy, and the Department of Health, Education and Welfare (which funds NIH).

Furthermore, not all requests for increased support for basic science were handled roughly by Congress. There was general support for new research in the energy sciences, for example. Many Congressmen have been impressed by the performance of Dr John Deutch as director of the Office of Energy Research.

Similarly the budget for the National Aeronautics and Space Administration emerged relatively unscathed. The two main casualties were the proposed search for extraterrestrial intelligence, and continuing research on lunar samples brought back by the Apollo missions; however major new starts, such as the solar–polar satellite and the earth radiation budget satellite, remained intact.

The two agencies to suffer most were the National Science Foundation and the Department of Defense. The former had been slated by President Carter for an increase in basic research funding of 9.7 per cent; it ended up with an increase estimated by NSF director Richard Atkinson as 5.6 per cent, well below the expected level of inflation (although still, according to NSF officials, better than some had feared.)

Finally Congress responded unsympathetically to administration requests for an increase in expenditure for military-related research. The Senate, for example, although indicating that it might be prepared to consider increases in existing university-based programmes, rejected a proposal to start a new "Defense Science and Engineering Programme" for university research which would have been allocated $9 million in 1979 and $27 million in FY 1980.

The administration had been hoping that some of the research items cut from the DoD budget could be included in a supplemental request now being prepared following a presidential veto of the final defense appropriations bill, which included provisions for building an extra aircraft carrier unwanted by President Carter. However both Senate and House committees are reported to be unsympathetic towards such a move, arguing that such requests should be placed in the 1980—rather than the 1979—budget.

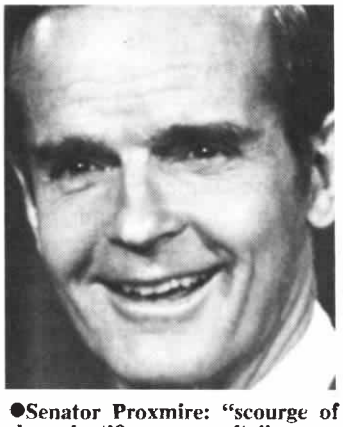
The net result of the various Congressional actions is what Dr Norman Hackerman, chairman of the National Science Board, has described as a "let-down feeling" among members of the scientific community.

Basic research, by agency ($millions)

	1979 Budget	Congress Approved	Change	% From Budget	% From 1978
NIH	856.0	1,019.0	+163.0	+19.0	+33.6
Other HEW	133.1	123.5	− 9.6	− 7.2	+28.1
NSF	754.9	741.9	− 13.0	− 1.7	+ 7.8
NASA	519.8	513.1	− 6.7	− 1.3	+ 9.6
Energy	467.5	465.5	− 2.0	− 0.4	+ 7.6
Defence	363.9	354.7	− 9.2	− 2.5	+10.8
Agriculture	251.9	256.2	+ 4.3	+ 1.7	+ 9.0
Interior	164.3	164.3	—	—	+ 4.6
Commerce	31.7	30.4	− 1.3	− 4.1	+10.9
Smithsonian	33.0	33.0	—	—	+ 5.8
EPA	30.0	24.0	− 6.0	−20.0	+14.2
Others	30.6	30.6	—	—	+10.1
Total basic research	$3,636.7	$3,756.2	$+119.5	+3.3%	+15.0%

From the AAAS report "Congressional Action on R&D in the FY 1979 budget".

●Representative Whitten: "no friend of basic science"



●Senator Proxmire: "scourge of the scientific community"

As far as next year is concerned, things look little brighter. The President has repeatedly stressed the importance he attaches to supporting basic science to guarantee the long-term health of the country. In his nation-wide anti-inflation speech of 24 October, for example, he said that despite other budgetary constraints, "federal support for research and development will continue to increase, especially for basic research".

Some observers in Washington feel that the budget request for 1980 will propose a growth rate for basic science of 2 to 3 per cent above the expected inflation level. They take comfort, for example, from the fact that one of the first visits paid by Mr Alfred Kahn, the President's chief inflation fighter, was on Dr Frank Press, director of the OSTP, to confirm that he agreed with Dr Press' emphasis on the importance of research and development for long-term national prosperity.

The great unknown is Congress. This month's elections sent a clear message to Washington: US electors are demanding greater prudence in the way that the federal government spends its money and popular support gathers rapidly for those seen willing to yield a budget-cutting knife.

As far as authorisation bills are concerned (those that represent policy decisions about the future directions of science) there is not likely to be too much problem. The new chairman of the House Science and Technology Committee will probably be Representative Don Fuqua of Florida, who has worked closely with agencies such as NASA in the past (and has several major research universities in his state).

Similarly the fortunes of science authorisation in the Senate are likely to remain in the relatively sympathetic hands of Senators Edward Kennedy and Adlai Stevenson Jr.

Appropriations committees, however —those that agree to disburse the funds —are a different story. In the House, the retirement of Representative George Mahon as chairman of the appropriations committee means that the position may be taken by Representative Jamie Whitten of Mississippi, the next in line on grounds of seniority.

Mr Whitten is no great friend of the basic science community, or of its peer review system. This year the appropriations subcommittee of which he was chairman rejected an administration request for $30 million for a competitive grants programme for agricultural research in the 1979 budget, an attempt to shift away from the present system by which research funds are handed out to state land-grant colleges on a pro rata basis.

Defending this cut—half of which was later restored in the conference with the Senate—Mr Whitten said: "Congress must not allow itself to be placed in the position of being held accountable to the people for their research priorities established by a non-elected bureaucrat issuing grants to his fellow scientists". And his subcommittee's report contained detailed directions on how agricultural research money should be allocated.

It is the Senate appropriations committee, however, which is giving the scientific community the greatest cause for concern, in particular over the implications of the loss of Republican Senator Edward Brooke.

The chairman of the subcommittee responsible for passing both NSF and NASA funds is Mr William Proxmire, scourge of the scientific community with his infamous "Golden Fleece" Award for esoteric sounding research. (It was Senator Proxmire, for example, who was responsible for deleting most of the lunar sample analysis from NASA's 1979 budget).

Proxmire strengthened

The results of the elections have strengthened Mr Proxmire's hand. Last week he said "The fiscal scalpels must be applied to both military and civilian programmes, foreign and domestic spending, and to the sacred cows of the powerful economic interests . . . No program, no interest group, and no region of the country, should remain immune to tough, detailed, and in depth budget cuts."

Up to now, Mr Proxmire's attempts to restrict funding for basic research has encountered strong—and frequently successful—opposition both from Mr Brooke, and from the ranking Republican member of the sub-committee, Senator Charles Mathias of Maryland.

However following Senator Brooke's departure, Senator Mathias is reported to be considering moving to a different subcommittee; and the concern of scientists that such a move could leave both NSF and NASA highly vulnerable has been heightened by the news that the subcommittee's minority staff member and aide to Senator Mathias, Mr Bob Clarke, is leaving the Senate to enter private law practice.

Much, of course, remains unknown about the future membership of committees, and will not be definitely resolved until Congress reconvenes in January. However one of the lessons brought home by the last Congress was the increasingly significant role played by intensive lobbying efforts.

Part of the substantial increase in support for basic biomedical sciences seems to have been due to a widely-publicised visit to Washington at the beginning of the year by a group of eminent biomedical scientists. In both public hearings and private meetings with Congressmen and their staffs, members of the group, which included three Nobel laureates and the heads of some of the leading US research institutions, argued that the long-term understanding of disease required a shift from a "disease of the month" mentality to a greater concern for basic science problems.

These arguments were subsequently reflected in Congressional action on the NIH budget. Funding for cancer research, traditionally a recipient of Congressional largesse, was increased by 6.8 per cent; in contrast, the research budget of the National Institute for General Medical Sciences was raised by 25 per cent, to a total of $231 million. Particularly significant was the emphasis placed in Congressional reports on the need to support "investigator initiated" projects, a point repeatedly stressed by the visitors to Washington, and the increase in funding for the Biomedical Research Support Grants programme.

A further illustration of the value of direct lobbying was memorandum sent by President Carter in June to the heads of appropriations committees and subcommittees requesting that they respect the integrity of the R & D package in his budget request. Although the success of the memorandum was limited, some feel that it helped limit damage to the budget of, for example, the NSF.

One message seems clear. Despite widespread awareness of the economic importance of research and development, the scientific community no longer enjoys any privileged position with respect to members of Congress, but must argue with everyone else for a share of the cake. "The universities, scientific societies and even industrial organisations will try to do even more than this year to increase support for research", Dr Jack Crowley, of the Association of American Universities, said last week.

It remains to be seen whether President Carter's concern for the health of basic science is any more successful next year than this. Certainly his public statements, and the close relationship which his scientific advisers have been able to build up with the Office of Management and Budget, have reassured the scientific community that, unlike some of his recent predecessors, they now have an ally in the White House.

But since much of the science budget is in the form of nonmandated expenditure (unlike, for example, social security and Medicare), it remains highly vulnerable to marauding budget cutters. And it will take more than fine words and noble sentiments to secure any real growth for basic science in the immediate future. □

Post-docs take up ARMS

Short term research contracts are leading scientists to rebel, writes **Joe Schwartz**

THE organisation of scientific research in the UK has evolved to the point where there are now 10,000 people employed to do research on short term contracts and an additional 1,800 PhDs registered as unemployed with the Department of Employment. Combine these figures with recent cuts in public spending and the result is the inevitable development of labour-management conflict in an area of the economy that has always considered itself immune from such pressures. And, wherever a management view in defence of short term contracts is articulated (contracts are for specific pieces of work, the scientific mind burns out early in life, unfruitful research lines can be easily terminated), there is a labour point of view articulated by short term contract workers themselves.

ARMS, the Association of Researchers in Medical Science, is one organisation that is attempting to speak for the needs of full time researchers employed in the nation's hospitals and medical schools on short term contracts.

Based at Guys Hospital in London, ARMS was formed after a series of public meetings in July 1978. Its principal goal is to establish a career structure for full time non-teaching researchers analogous to university lectureship scales. Although not ruling out trade union affiliation, the organisation at present intends to negotiate with employers and to do educational work in an effort to change the present arrangements.

Membership in the organisation received a boost in September when the University of London which administers Guys instituted waiver clauses in short term contracts. Seeking to prevent 'onerous benefactions' to the research staff, the university demanded that researchers relinquish their rights to redundancy pay and claims against unfair dismissal. The outcry against this particular action has cast the university into the role of a 'bad' employer and could be rescinded if the researchers are able to maintain an active resistance to the measure. Brian Davis, Labour MP for Enfield and a member of the MRC, acknowledges that the MRC would like 'maximum flexibility' in its hiring and firing but is working to abolish the waiver clause as soon as possible. The Cancer Research Campaign, located at Mt Vernon Hospital, has removed the waiver clause after protests from the researchers there.

The waiver clause, however, is only the most recent grievance. Dr Mick Kadlubowski, Information Officer of ARMS, regards administrators' arguments in favour of contract work as specious and self serving. "The system prevents the untenured full-time researcher from having an effective say in research policy. Should a project be successful the honours accrue to the grant holder; if not, then support for the unfruitful line can be terminated and the untenured scientist is once again looking for a job". This practice particularly affects women scientists. It is common for younger women scientists to be given the riskiest research lines since senior researchers tend to take their scientific ambitions less seriously than those of their male colleagues because women bear children.

As for the question of 'burn outs', Kadlubowski points out that such a fate does not seem to affect administrators and clinicians. "Project grant applications are on the whole successful only when submitted by those with tenure and are usually assessed by their tenured colleagues. The ability of these senior academics and clinicians to devise and judge a research proposal would not appear to wane".

According to ARMS' analysis the present system is costly and wasteful.

One half of a researcher's time on a short term contract is spent settling into the project and then towards the end looking for a new job. After two or three contracts, the researcher is 30–35 years old and the contractors begin to find the price of experience too great. The experienced researcher is told to seek employment elsewhere and the funds thus 'saved' are used to boost the intake of juniors. "How can science progress when hard-won experience is so callously thrown away", asks Kadlubowski.

Davison of ASTMS; Akker of AUT.

ARMS has approached the DES and the MRC with little effect. Dr Anne Simmonds, chairperson of ARMS, is scathing about a recent MRC grant scheme: five-year renewable grants to age 35, grants to permit tenured academics to do full time research for a short time, and external full time research professorships for the clinically qualified. "At this critical time with their scant funds and these alarming unemployment statistics," she says "the MRC comes up with funds to allow the tenured time off to dabble in full time research, the clinician the opportunity to return to the field and progress without restraint to the professorial level, and another ten years of insecurity offered to the non-medically qualified."

The action of the MRC in totally ignoring the most urgent problems, the scandalous plight of full-time non-medically qualified researchers despite representations from clinical professors, deans of medical schools and the scientists themselves is utterly reprehensible".

Anne Simmonds is equally critical of university administrators. "Instead of raising their voices in support of some urgent action to resolve the situation they are aggravating it by looking for loopholes in the law to enable them to escape any commitment to their full-time researchers whatsoever."

Faced with this response from employers ARMS members have contacted ASTMS and the AUT, the two trade unions involved. "ASTMS calls us elitist and they want money on the line before they'll do anything", say members of the executive committee. Stan Davison, assistant general secretary of the 460,000-member ASTMS, denies this charge but admits that the problems of organising short term contract workers are formidable; the workforce is spread out over the entire country in small pockets of twos and threes; many scientific workers are indifferent or even hostile to trade unions; and research workers tend to be inexperienced and sympathetic to management arguments.

ASTMS is opposed to short term contracts and the unregulated em-

Dr Joseph Schwartz is on leave from City University, New York.

0028–0836/78/0276—0310$01.00

© Macmillan Journals Ltd 1978

ployer-employee relationship that currently prevails. Philip Hounsell, division organiser for ASTMS at Guys Hospital welcomes an organisation like ARMS. Hounsell feels that only by rank and file organising at the local level can sufficient strength be built to have a real effect. Both Davison and Hounsell emphasise however that without proper trade union organisation and support it is doubtful whether the authorities can be persuaded to act.

The 30,000 member AUT is approaching the problem at a higher level. It has negotiating rights for the universities and successfully negotiated national salary scales for researchers in 1975. John Akker, deputy general secretary, says that the AUT handles "scores of cases" of senior research workers, many with international reputations who have been made redundant at the age 40–45. The hardship is even worse he says because "many people are so specialised as to be unemployable".

The AUT offers legal services to members faced with dismissal and has been able to negotiate cash settlements in many cases and is in the process of formulating a code of practice for the employment of research staff. The AUT is attempting to move its efforts into the ministerial arena. "The MRC claims that the universities are the employer and the universities say since the MRC provides the money it is up to the MRC to sort it out", says Akker. Even if the research councils were to become actively involved there are still 78 autonomous institutions to deal with. The AUT wants the government to convene a committee with trade unions, universities, research councils and government representatives on it.

Whether the government is prepared to take on the issue is unclear. Certainly the universities are acting as if they have just discovered 19th century labour relations. Insisting on the value of piece work places academia in the league of garment manufacturers in a pre-trade union situation. This sensibility is reproduced in many researchers themselves who accept the line that a good student can always find a position and who reject trade union protection as being unprofessional.

But since research is an integral part of a modern economy some pressure exists for researchers to be able to make some changes. Certainly young potential scientists will begin to reject a scientific career if they become aware of the true conditions of research work. And there is nothing to prevent out of work scientists from organising just such a campaign if they are angry enough. □

People before pounds, say SRC

BRITAIN'S Science Research Council is no longer so worried about money, but is finding itself with another problem: manpower. The chairman of the SRC, Professor Geoffrey Allen, said last week that "the manpower problem is preventing us from diversifying".

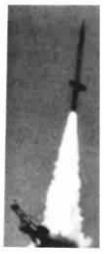
Professor Allen was presenting a fairly encouraging annual report for the year to April 1978. The budget reduction of 1.7% per year imposed on the SRC by the Advisory Board for the Research Councils had been held to 1%, and merely another £20 million over the next 4-5 years would make the SRC comfortable. A 2-3% increase in the 'science vote' (the money allocated to the ABRC by the government for disbursement to the research councils) would give the ABRC all that the ABRC thinks is needed nationally.

Professor Allen is taking the training function of the SRC very seriously, and hopes to introduce courses for "topping up" in new technologies. This is "the major issue now" in training. And he wants to introduce a five person "think tank" into the Council which will concern itself with improving the links between university and industry. This call for five new people underlines Professor Allen's concern for SRC manpower.

The development of new activities at the Rutherford and Daresbury laboratories, whose subnuclear physics accelerators were closed in the year under review, is also raising manpower problems. The SRC, like many other government bodies, is under a directive to reduce its manpower —in the SRC's case at 1% per annum. This would mean a loss of some 30 people per year; but with fairly young staff it is a hard target to reach through natural wastage and early retirement.

While it has been possible to retrain many of the displaced high energy physicists at Rutherford and Daresbury in subjects like energy research and computing, it is not so easy to turn a physicist into a biologist to run, say, a biology support team at the new synchrotron radiation source. The central problem is that a few key people are needed, with special talents, to implement the changes which Professor Allen would like the SRC to undergo. "We now need different types of scientists to maintain our central facilities and links with universities", says Professor Allen. The SRC is a potential employer of scientists. But at the moment it is being held back.

Robert Walgate



Skylark rockets could be made from spare parts say UK geophysicists

A GROUP of UK geophysicists has come up with an idea which, in the early 1980s, could help soften the blow on auroral physics of the Science Research Council's (SRC) cutback in 'big science'. It has suggested that component parts left over from the SRC's abandoned Skylark rocket programme could be used to build as many as twelve extra Skylark rockets.

The decision to abandon the Skylark rocket programme was taken in 1975-6 as part of the SRC's plan to economise on 'big science'. Since then the British Aircraft Corporation, the rockets' manufacturer, has been slowly winding down production. Only one rocket now remains to be launched from Woomera, Australia, next May. During the 21 years of the programme, more than 250 Skylark rockets have been successfully launched.

The UK rocket programme now depends on two small rockets, the 7.5 inch-diameter Petrel rocket capable of launching 14 kg to a height of 200 km and the 10 inch diameter Fulmar rocket capable of launching 50 kg to more than 250 km. These are used for studying the lower and mid-ionosphere.

The 17 inch-diameter Skylark rocket is capable of carrying about 150 kg to altitudes as high as 750 km. It is the only UK sounding rocket developed so far which is capable of investigating the structure of the upper ionosphere and the origin of auroral particles. Although a Fulmar might have the ability to fly high enough, it does not have the capacity to carry the instrumentation needed to measure energetic incoming particles and electric fields.

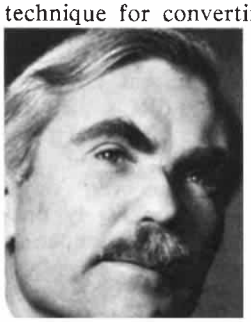
The geophysicists claim that twelve Skylarks of varying size could be assembled, furnished with payloads and launched at no extra cost to the SRC's planned rocket programme over the next ten years if they were launched in fours at three-year intervals. This would mean launching fewer Petrels and Fulmars; but the value of the Skylark is such that it is thought there will be little opposition. Just how much support the idea could gain is what the group of geophysicists is trying to find out at the moment. By late next Spring it hopes to have sufficient knowledge of the interest and the kinds of experiments which could be flown to put forward a concrete proposal to the SRC for funding.

Judy Redfearn

© Macmillan Journals Ltd 1978

● **Nuclear waste disposal:** Highly radioactive waste from Britain's nuclear reactors could be safely stored underground by the 1990s, if suitable geological sites could be found, according to Dr Lewis Roberts (pictured below), director of Harwell research laboratory. In a lecture to the British Nuclear Society recently, he reviewed the present policy of the UK Atomic Energy Authority on radioactive waste disposal.

The means of solidifying high level wastes produced by the reprocessing of nuclear fuels, and the technology for building repositories, should soon be available, he said. At the moment, waste was stored as a highly radioactive acid solution in large tanks at Windscale. However, a solid was easier to contain and control than a corrosive liquid, so a technique for converting waste into a solid glass composition was developed at Harwell. This "vitrified" waste could then be placed in an underground store fitted with a simple cooling system and left, with minimum human surveillance, for the 500 years it takes for the highly radioactive actinides to decay.



An international programme, backed by the EEC, was underway to assess different sites for final storage, Dr Roberts said. Germany, Italy, Holland and Belgium were investigating the potential of salt and clay deposits, while Britain and France looked at hard crystalline rocks such as granite.

British exploration were running into trouble because some planning authorities (for example, in south-west Scotland and Northumberland) refused to allow the UKAEA to drill test holes. More research was needed to determine how safe the repositories would be, and the behaviour of containers and glasses under the conditions which might occur in storage sites. Despite these set-backs, Dr Roberts believed that research into final storage of radioactive waste should continue. The UKAEA was also looking into alternative storage sites, such as the oceans.

● **MPs seek comments on genetic engineering:** The social and political implications of the use of techniques of genetic manipulation, both in research laboratories and for industrial processes, are to be investigated in the UK by the Genetic Engineering Sub-Committee, which was set up by the House of Commons Select Committee on Science and Technology during the last session of Parliament.

The committee is asking any interested people or groups to submit written reports, even if they have not been specifically invited. The committee has also approached Government departments and agencies, professional bodies, companies and trade unions for reports on a wide variety of technical, commercial and ethical considerations. Submissions should be made before the end of the year, and the committee expects to report some time in 1979.

● **Chemical company claims dioxins a natural by-product of combustion:** The Dow Chemical Company, defending itself against charges that traces of chlorinated dioxins—one of the most poisonous of industrial chemicals—found in fish in the Tittabawasee River in Michigan had leaked from the company's pesticide plant, claim that its scientists have discovered the compounds to be natural by-products of combustion.

In what is said to be a "significant breakthrough" in the chemistry of fire, the company scientists say that small quantities of chlorinated dioxins, well below the level that poses a hazard to human health, have been discovered in coal and oil-burning power plants, in home fireplaces and in charcoal grills. "We now think dioxins have been with us since the advent of fire. The only thing that's different is our new-found ability to detect them", Dr Robert Bumb, research director of Dow's Michigan division, said last week.

The eating of fish from several rivers near the plant was banned by state and federal environmental agencies earlier this year following the discovery of traces of chlorinated dioxins in both soil samples and game fish.

● **Petrol additive found to be potent carcinogen:** A chemical widely used as a petrol additive and a fumigant for agricultural crops has been found to cause a "high incidence" of cancer in laboratory animals by the National Cancer Institute. According to the institute, tests have shown ethylene dibromide (also known as EDB or 1,2-dibromoethane) to be one of the most potent animal carcinogens ever found, and must be considered capable of causing cancer in humans.

The National Institute for Occupational Safety and Health has estimated that in 1977 about 9,000 workers were exposed to EDB in industrial settings, and that a further 650,000 service station attendants were exposed to much lower levels of EDB. The chemical is added to tetra-alkyl anti-knock fluids to remove lead from the engine's combustion chamber.

In a two-year study, during which large doses of EDB were administered to rats and mice, stomach cancers developed in 58% to 90% of the rats and 56% to 94% of the mice, depending on the sex of the animal and the dose received. Some cancers developed less than two months after the beginning of the experiment, and the death rate among experimental animals was so high that the experiments were discontinued when the programme was only half-way through.

● **World population growth is slowing:** The rate at which the world population is growing has begun to slow down. Since the 1960s, birthrates have fallen by 15% in about half of the countries of the developing world, according to Rafael Salas, executive director of the UN Fund for Population Activities. However, Salas adds: "It would be unwise to conclude that the population problems of individual nations as a whole are at an end, or anywhere near it."

Another big world population rise is forecast, since at least 40% of the people of the Third World are under 15 and approaching child bearing age. In his annual report, Salas puts the lowest estimate for world population in the year 2,000 at 5.8 billion compared with today's total of about 4 billions. In poorer countries the age structure is changing towards a higher proportion of very young and very old people. The trend is causing concern at the UN, because it will put strain on economic and social resources for education and employment.

● **Naming the unifying gauge theories:** Some physicists are suggesting that the time has come to rename some of the new forces and interactions in particle physics. Greek, which was the basis for nomenclature until the arrival of modern idiosyncrasies such as "colour" and "flavour", could be used as the source for the new words, some physicists suggest.

The theoretical developments which appear to have unified the weak and electromagnetic interactions failed to

© Macmillan Journals Ltd 1978

provide a new word that physicists could use to describe the forces, apart from the indirect term the "Weinberg-Salam model" and the unimaginative "electro-weak". A candidate for the title is "asthenic force", originally suggested by Chanowitz and Ellis and now revived in the October edition of the CERN Courier. *Asthenis* is a Greek adjective meaning "not strong", and so the word has the advantage of covering all phenomena not associated with the strong force. Chanowitz and Ellis suggested that the unified gauge theory of weak and electromagnetic interactions could be described as "quantum asthenodynamics". Pursuing their Greek source further, they proposed that the grand unified gauge theory of all the forces could be named "quantum holodynamics".

● **Lasers:** The laser light show in London's Oxford Street, the first of its kind to be used in Christmas decorations, was switched on last week, despite a safety inquiry which nearly banned the show. The six krypton and argon lasers have splitters to produce red and green beams of light. The system was originally built to sweep the laser beams along Oxford Street 20 feet above the ground. But the risk of a beam being reflected by a window into someone's eyes forced the Health and Safety Executive and Westminster Council to insist on the lasers being fixed in one position. The beams would be allowed to move only if every window along the route was blacked out.

'Problem' solving still claims Poland's scientists

"DESPITE many shortcomings which still exist, we have the right to state that our science is carrying out its tasks more successfully each year". So claimed Piotr Jaroszewicz, Prime Minister of Poland, speaking at the opening session of the Sejm (Parliament) last month.

For a head of state to present the Government's programme on science policy is somewhat unusual. Partly it may be attributed to what Jaroszewicz called the "particularly active part" which science must play in overcoming "the difficulties experienced by our economy at present" (disastrous results in agriculture for the fifth year running, difficulties in earning the hard currency needed for vital raw materials and over-consumption of labour-time, energy and raw materials).

Moreover, in September, Poland's trouble-shooting Minister of Science and Higher Education, Sylwester Kaliski, died after a car crash, and the presentation of the science programme by so eminent a figure as Jaroszewicz seems a further assurance of government commitment to the policy so largely associated with Kaliski's name, although originally worked out by the Polish Academy of Sciences —a structured hierarchy of R and D "problems".

In this scheme, problems vital to the economy range from "State problems", through "Key", "Interdepartmental" and "Departmental" down to "Branch" problems. The differences are both administrative (who is responsible for funding and coordination) and a matter of actual priority.

There are seven "State" problems at present—copper, coal, building industry, food, electronics, cancer and water/environment, and it seems likely that, in the near future, energy and the chemical industry will receive the same status.

The urgency of the "state" problems is considerable. In the case of building, the construction of a new campus for Poznan University has been approved and introduced into the central plan of the Government, with allocation of 4,500 million zloty. Yet, according to the pro-rector, Dr S. Kozarski, the University is having considerable difficulties in getting the construction under way, simply because of the overall state of the building industry.

According to a spokesman for the Ministry of Higher Education and Science, the system of problems is "working, though not so well as we had hoped at first". During the next five-year plan, he said, there would be greater freedom of decision-making left to the researchers and fewer referrals to distant committees in Warsaw. At present, the heads of research teams are far to circumscribed by "reports and byelaws".

Does this direction from above restrict academic freedom? No, said the rector of the Poznan Polytechnic, Dr Boleslaw Wojciekowicz, where some 60% of research comes under the central scheme. There is no compulsion, he said, for any institution to accept any problem it does not feel it can tackle. "Freedom of science", the rector exclaimed, "includes the taking on or not taking on of government research".

The system does serve as an indicator to research institutions what is important now and in the future, and provides a method of concentrating both personnel and financial resources. Moreover, any institution—his own Polytechnic, for example, can at any time propose some new problem to be included in the scheme. At present, he added, his Polytechnic acted as research coordinator for three such problems, at the "Interdepartmental" level, coming therefore, under the joint authority of the Academy of Sciences and the Ministry.

There also remains the other 40% of research activity at Poznan outside the central scheme. Some of this includes direct contract work for industry, but funds are still available for basic research direct from the Ministry of Education, and are distributed to the individual Institutes concerned. The Directors of these institutes have a fairly free hand in allotting these funds, and can finance any project they consider promising.

So far, too, international cooperation does not seem limited to the formal Government policy of Comecon cooperation. According to the Ministry spokesman, since Comecon is primarily an organisation for economic matters, the intra-Comecon links in higher education are somewhat weak. When there is joint research, it is directed not through the Ministry of Education but through the specialised ministries which have their own research institutes at their disposal.

Dr Wojchiekowicz put it more bluntly. "If the ministry proposes a key problem to us, we are not interested in whether another country needs it as well—though of course if we know others are working on it, we naturally try to establish contacts". His Polytechnic, he said, had collaboration agreements with 26 foreign institutes—including some in West Germany, Holland, France, and the USA. "But with non-Comecon countries, the basic problem is the perennial one of currency."

It would be strange, indeed, if the precious funding for foreign cooperation did not go first and foremost to government proposed problems, and, indeed, a number of examples could be cited. There is close cooperation with the UK on safety in coal mines, for example (coal is a "State" problem), while the "urban forest" conservation scheme (a "Key" problem) benefits from Landsat photographs under a

© Macmillan Journals Ltd 1978

NASA–Polish agreement.

It would seem inevitable that this concentration on the coordinated problem concept should have its effect on the higher educational system generally. Jaroszewicz's policy speech proposed that "scientific and research topics of particular importance of the economy should be stressed even more not only at the research stage but in higher education also."

Likewise, the Ministry of Education spokesman indicated that future plans include "establishing closer links during study with productive activity". He also said that serious consideration was being given to the possibility of "reattestation"—the granting of professional qualifications not for life but for a limited period only, with renewal depending on the candidate's career in the meantime.

Not surprisingly, perhaps, a number of academics are privately concerned about the ultimate effect of the problem concept on Polish science. The freedom to accept or reject a government proposal is, they feel, a somewhat circumscribed one. Academic freedom, they point out, was one of the principal issues of protest against the proposed constitutional changes in 1975-76. Although the ensuing "debate" resulted in a certain toning down of the clauses relating to Party preeminence, there still seems to be considerable nostalgia in certain circles for the concept of autonomous universities and academies electing their own senior officials without outside influence, instead of requiring, as at present, government and Party approval. **Vera Rich**

Leather science as alternative technology

by Paul Calvert
who is on leave from the University of Sussex

In developed countries it is easy, as a scientist, to adopt the view that most research stands a chance of being applicable and most applications of science will lead to an improvement in our overall well being. This is much less true in developing countries; given their labour surplus and shortage of capital any technological advance which replaces men with machines may be bad. An example of a research organisation trying to serve a traditional, cottage industry is the Central Leather Research Institute (CLRI) in Madras.

The CLRI was set up under CSIR in 1952 with the task of providing scientific support to an industry which consists mainly of small factories and individual, often illiterate, craftsmen spread throughout India. An additional, uniquely Indian, dimension to the problem is that most of the workers are members of the "untouchable" castes. Until 1971 the director of the institute was Professor Y. Nayudamma who subsequently became director-general of CSIR.

Writing in 1967 (*Science and the Human Condition in India and Pakistan*, edited by Ward Morehouse, Rockerfeller University Press, 1968) Nayudamma described the philosophy and operation of the CLRI. Much of the institute's effort is put into extension centres close to concentrations of the leather industry and extension workers who can visit and live in the villages. These workers must respect the skill of the village craftsmen and must be able to talk to them at a level they can understand. Thus a method for controlling *p*H during tanning would be presented simply as a series of coloured packets to be added at different times in the process.

Other programmes include a "guest tanner" who can visit and work in the institute, and monthly visits by the director, with local officials in tow, to the industry in a particular state. Institute staff have also been sent in to revive businesses which are running at a loss, the idea being that this can be a powerful demonstration of the benefits of science. CLRI is closely associated with the University of Madras and runs BSc and PhD courses in leather science as well as refresher courses for teachers from regional tanning institutes.

Since going to CSIR, Nayudamma has been able to develop his ideas of the role of science in India and has propounded them in two recent lectures. In a paper called "Endogenous development: science and technology" given at a UNESCO meeting in Paris in April, he discussed science and technology in developing countries. In his eyes development is not simply a question of economic growth but of development of a people in terms of dignity, resourcefulness and social justice. Such patterns for development cannot be imported since they must suit the society in question.

Nayudamma quotes Gandhi's vision of self reliant, egalitarian rural communities using labour intensive methods and local resources but intellectually open and active. Science, based as it is on experiment and reason, should reduce the hold of superstition and increase rationality and enterprise.

Nayudamma sees imported technology as imposing the values and culture of the developed countries on the less developed. The success with which Japan imported technology is due to the strength of their indigenous technology which enabled them to modify the imported methods to suit their own needs. So some non-imported route to technical development is needed in order that a level is reached at which constructive use may be made of imported technology. This "alternative technology" has as its goals to provide employment, meet basic human needs and build confidence and self-reliance. It would include use of local resources, recycling, have a low energy demand and would be a weave of traditional and modern methods.

At the moment manufacturers tend to prefer imported processes which come fully developed, with known brand names as compared with local inventions which are untried and involve more risk. To overcome this, processes are offered with the government providing 50% of the risk capital and guaranteeing the result.

In "Developing Science for Development" a lecture at IIT in Madras last year, Nayudamma outlined the kind of research organisation he saw as necessary to serve a developing country. The key to science within this framework is one of setting clear goals for researchers. At present they tend to be judged only on their status in international science so there is little encouragement for working on strictly local problems, particularly for scientists trained abroad.

Further, scientists must be in close contact with the rural population; it is important to know the people whose problems you are solving and these are not the westernised elite but the rural majority with more basic needs. This lecture was given at a time when CSIR was under threat of being broken up and apportioned amongst the various ministries, a good example, perhaps, of imported ideas (see *Nature* **270**, 89; 1977). Nayudamma strongly defends the existing structure which survived, and calls for a clear lead from the government as to what technology is needed. When clear objectives are defined scientists usually achieve them.

In helping the leather industry CLRI must also tend to come into conflict with the caste system which is deeply ingrained in rural society. In recent years the untouchable castes including leather workers have tended to move from the villages into the cities where they have more political and social freedom. In the case of leather workers this will include setting up small businesses or co-operatives. By becoming involved with small scale industry,

© Macmillan Journals Ltd 1978

scientists thus also become very exposed to the ebb and flow of political forces.

Current fundamental research at CLRI, in addition to studies of tanning and degradation in leather, include extensive work on grafting synthetic polymers to collagen (*Makromol. Chem.* **175**, 729 (1974); *J. Polymer Sci.* **9**, 3199 (1971); *J. Appl. Polymer Sci.* **16**, 1975 (1972)) to produce wear resistance, water repellant and glossy surface finishes on leather. This sort of work is coming more into its own as India exports more finished as opposed to semi-finished leather. These treatments can be simply carried out and so are appropriate to small-scale industry. The institute also provides courses in design of leather goods so that the industry is able to keep up with fashions in world markets.

Thus CLRI provides an example of a scientific institute serving "alternative technology" and demonstrates the considerable effort that must be put into keeping communication open between the users and the scientists. Nayudamma, following his own precept that a director should return to the laboratory at the end of his term, is back at CLRI as a "distinguished scientist". He is studying the impact of CLRI on the leather industry over the last 25 years. The current director is M. Santappa who is also professor and head of the department of physical chemistry at the University of Madras. Santappa's main scientific interests before becoming completely involved at CLRI were in polymerisation of synthetic polymers.

The big question confronting this approach to science is whether a technically developing cottage industry could ever compete with heavily capitalised manufacturing industry. About 10 years ago much effort, particularly by Du Pont, was put into making artificial leathers consisting basically of a foam and skin structure of PVC. Had this succeeded, as it almost did, it would have drastically reduced the market for natural leather. In an industrialised society if workers lose their jobs under these circumstances they have a reasonable chance of getting others. Traditional craftsmen, such as Indian leather workers, do not have this flexibility, particularly if they are modernised to the extent that they depend on exported goods for their livelihood. □

Third World needs 'barefoot' science

How does a water tap work? **Michael Moravcsik** and **R. H. B. Exell** explain why knowing the answer can be a breakthrough for an African village.

The practical need for science in developing countries is rarely discussed. Debate usually rises to a general, analytical, or theoretical level. There is no question that such discussions help in understanding the role of science in those countries, and assist the proponents in convincing incredulous officials in national, international, and other decision making organisations, that science is necessary. And they also boost the morale of the scientists in the developing countries—who otherwise have many reasons for being discouraged and frustrated. But one should also be aware that the need for science can be much more down-to-earth.

In this brief note, therefore, we want to present a few examples where exposure to science as a way of thinking and as a method of inquiry and problem solving could play an important role in very simple everyday situations.

The examples are taken from "real life" situations that we have encountered during our stays in developing countries. Colleagues of ours could doubtless add dozens of such examples to the list.

Michael Moravcsik is from the Institute of Theoretical Science, University of Oregon, USA; R. H. B. Exell is from the Asian Institute of Technology, Bangkok, Thailand.

Keeping cool: In urban buildings in tropical areas, built using modern techniques, the top floor is often unbearably hot. The reason is simple: the usually flat roof absorbs solar radiation, and no provisions are made to either insulate the roof from the ceiling of the top floor, or to design the building with a simple double roof with air space between the two for ventilation. Ironically, such features are often included in the traditional, older, and rural structures, but since the scientific reason for such features is not comprehended, builders cannot distinguish between those features of traditional architecture which are functional and those which are ceremonial or due to limitations of construction technology.

Keeping hot: In numerous developing countries tin-box ovens used on charcoal stoves are common. They simply consist of a usual thin-walled surplus tin-box, placed on top of a regular iron charcoal stove (similar to the hibachis popular in Western countries for use in backyard barbecues).

It does not take very much science to realize that the efficiency of such an oven, and also the attainable temperature inside, could be increased if all sides of the tin box except the one in contact with the stove were insulated. A little more science tells us that such a saving in fuel would be quite substantial, very rapidly returning the small initial investment in one of the available insulating materials.

Current problems: Electrical wiring in developing countries is often installed very flimsily and haphazardly, so that breaks in the wiring and other types of damage occur frequently. In such instances there are many accidents as well as further damage due to incorrect rewiring. For example, we have seen instances when the break in a two-strand wire was "repaired" by joining all four dangling wires in one knot.

Even the most qualitative and simplest *conceptual* understanding of current as a flow (and hence a schematic knowledge of what bears the unnecessarily forbidding title of Kirchhoff's laws) would do much toward preventing such occurrences.

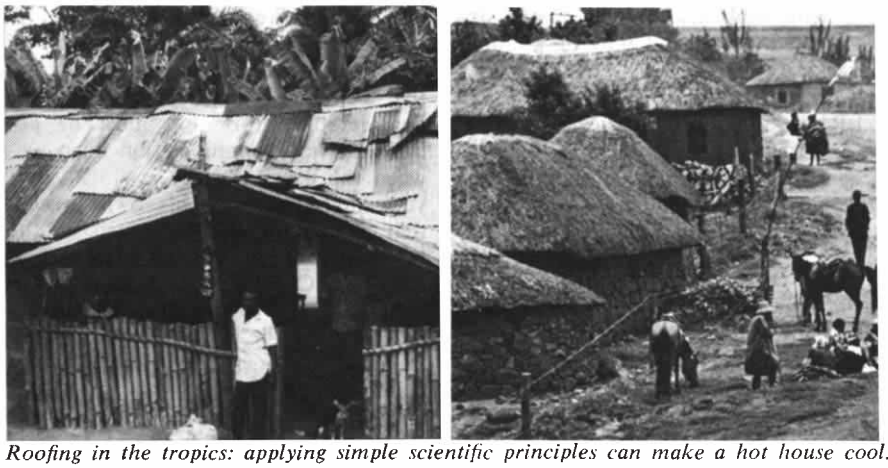
Turning on: Recently, in one of the developing countries, the supply of tap water became quite intermittent for several months. It was then found that during those hours of the day when the water supply did not operate, much subsequent loss of water was initiated by taps being left in an "on" position during the time when the supply was off. People would try the tap, see that no water was coming, and then leave the tap without turning it to the "off" position. When the water supply re-

0028–0836/78/0276—0315$01.00

© Macmillan Journals Ltd 1978

turned, the water kept flowing sometimes during the whole day until somebody happened to return and turn it off.

Water being expensive, one of the important contributing factors in this apparent neglect is likely to be the complete absence of an image of a water flow with a tap being a gadget to transmit or obstruct this flow. When this simple point is not understood, there is no more "reason" for turning the tap to the "off" position when the water fails to emerge. This is science on such an elementary level that most of us might find it difficult even to realise that it is science.



Roofing in the tropics: applying simple scientific principles can make a hot house cool.

Counting squares: The next two examples are perhaps more mathematical than physical, but they again stress the importance of understanding concepts rather than following recipes.

A junior lecturer at a university, engaged in an "applied" problem in botany, was confronted with the task of measuring the surface area of leaves. He kept postponing the measurements, however, because an expensive planimeter ordered from abroad had not arrived and, according to some textbook, was *the* instrument to use for such measurements. He was amazed that tracing the leaf on to readily available when one of us pointed out to him able graph paper with a square millimetre network and then simply counting the squares within the traced figure yields the desired answer with as great (or greater) precision than would have been obtained with the planimeter.

Saving time: In mathematical calculations related to applied scientific or engineering problems, the lack of a conceptual understanding of mathematics often prevents the practitioners from using approximation methods (particularly numerical ones) which yield the answer to the required degree of precision much easier than fancy analytical methods or computer evaluations of the "exact" problem. When in a calculus of variations problem, the shape of a curve has to be optimised, the approximation of the smooth curve by a chain of straight-line segments can reduce the problem to the solution of a simple set of algebraic equations.

The same is true in complicated boundary value problems in differential equations of heat transfer and other problems. Countless hours of precious computer time could be saved on expensive and scarce electronic computers in developing countries if their users were better trained in the fundamental concepts of science and mathematics and relied less on recipies from engineering handbooks or on imitating calculations described in esoteric books dealing with different problems and different tolerance requirements.

Hot water: A solar water heater is to be designed. Even if the basic qualitative principles of such a design are understood, much will depend on the ability to carry out a fairly reliable estimate of heat losses, conversion efficiencies, and the like. For this the scientific principles underlying these phenomena must be clearly understood, because engineering handbooks or the lecture notes saved from university training will hardly give automatic recipes for such "non-standard" situations.

If an engineer claims that he could make steam with such a solar heater, he must also be able to estimate with sufficient accuracy the temperature and the amount of such steam. In the absence of such ability, his project remains an academic exercise and cannot be introduced to "real life" situations where the economic considerations are important and where a quantitative assessment is essential.

A similar situation exists when a claim is made for a design of a practical solar-heat dryer of agricultural products. In this case the useability of the resulting products as well as their quantity will depend crucially on the quantitative factors in the design which, as before, must be estimated on the basis of the scientific principles underlying the engineering applications.

So much for the examples. In seeing them, the reader may advance various objections, at least two of which should be briefly discussed.

The first objection may be that our examples depend not so much on the lack of science but on the lack of "common sense". It is relatively easy to dispose of this objection. Indeed, to a person imbued with some conceptual background in science and technology, the examples may appear a matter of common sense. This is so, however, because he has internalised simple concepts of science and mathematics to such an extent that viewing the world in that light now appears to him a matter of common sense and no longer a matter of science and mathematics.

This is the main point we wish to make: in developing countries such internalisation and fusion of scientific and mathematical conceptual elements with the customary empirical view of the world is very rare, whereas it is much more frequent (though far from universal) in the more advanced countries. This shaping of man's world view is, in fact, one of the primary goals of science in developing countries, and yet it is one that is much too infrequently included in the discussions that prevail throughout "development" literature.

The second objection may be that the missing elements in each of our examples could be provided by a well-trained technologist and engineer and hence science is not really directly involved. In some idealised world this statement may carry some validity. In practice, however, the technological and engineering education, particularly in the developing countries, is such that a functional understanding of the conceptual scientific basis of technology is extremely rare among the graduates of technological and engineering institutions. This is partly so due to directives based on a total misunderstanding of the nature of applied scientific research and of technological development.

According to such directives, engineers and technologists should be given a narrow education in which "unnecessary" abstractions are weeded out and training with respect to a very narrow set of specific projects is the goal. In part, however, the lack of conceptual understanding is due to the lack of contact of these students with teachers who are, themselves, involved in some scientific work and who therefore have had a personal experience with how science is actually done, as distinct from how it appears from the textbooks. Thus we conclude that we need a much greater attention to education.

correspondence

The Natural History Museum needs the best of both worlds

SIR,—The current controversy concerning the Natural History Museum's new exhibition scheme and philosophy, eptiomised in the opposing views expressed by Drs Miles and Halstead (26 October, page 683), highlights problems which certainly deserve the "fullest possible public discussion" as Halstead suggested. While my personal opinion is no more (or less) valuable than that of the next man, as the father of two of the museum's avid users, and also as an ex-curator of one of its scientific sections, I can, at least, see the situation from two different points of view and can make some observations rather more frankly now than when I was still a staff member.

The tremendous gulf between the expressed views suggests that the fundamental question of the museum's role remains to be resolved—despite two centuries of more or less continuous discussion. No-one would dispute the fact that the museum has two distinct general functions, both clearly identified, as Miles points out, by the British Museum Bill of 1753, and as valid today as they were then. The first of these, to carry out research into the natural world, occupies almost two-thirds of the museum's total complement of 750 and sails serenely on with hardly a ripple from the outside world to rock the scientific boat. It is the second, more public role, which attracts all the attention even though it is dealt with on only seven sides of the museum's most recent 170 page report to Parliament.

But while these general functions have remained unaltered, the techniques and approaches needed to discharge them have changed, and will continue to change, throughout the museum's existence. For just as it would be totally inappropriate for the present scientific staff to approach their work in the same manner as their Georgian predecessors, the public galleries should surely bear little resemblance to those of 200 years ago. For one thing, both in numbers and in social stratum, the public served then was very different from that using the museum today. In its early years no more than a few dozen privileged visitors could view the British Museum exhibits each day—and even then only by appointment with a Keeper; as a son of the eighteenth century equivalent of a lorry driver I would personally have had little chance of stepping through the door, let alone of becoming a Curator!

Second, the museum now operates in a different educational environment. The present-day flood of scientific information for the "layman" through books, films, radio and, above all, television did not exist ten, or even one, generations ago. No museum exhibit should attempt to compete with the slickness and immediacy of a television presentation, for whatever you may think of their particular contributions, a Bellamy or a Cousteau can reach a larger and more varied audience in one hour than the BM(NH) could hope to reach in ten years. Surely the museum should adopt a complementary role in these circumstances, concentrating its efforts on those things which it can do better than the mass media by exploiting its particular strengths, its in-house scientific know-how and its vast collections of material.

The official museum view of the way forward seems to be to attempt to put over concepts and ideas to a general public largely devoid of any background in natural history. The opposing "Halsteadian" view, which appears to have the support of a significant proportion of the museum's own scientific staff, seems to favour a return to a more traditional type of "scholarly" display with many examples of the animals, plants and minerals themselves from which the informed visitor can draw his own conclusions. Either approach alone would be dangerous, sterile and damaging to an institution for which so many of us feel the deepest affection. If Dr Miles really believes his survey results which told him that the old displays taught virtually nothing, he should have been with my wife and I when we took our 6-year-old son on his fifth or sixth visit to the museum some years ago. After a depressing appointment with a child psychiatrist had confirmed that he had severe learning difficulties, we were considerably encouraged to have him point out to us the differences between African and Indian elephants in the now dismantled Main Hall exhibit! Like most kids he still likes to see lots of dinosaurs, whales, fishes and so on, and much prefers them to the Hall of Human Biology, which, after its first month, was looking pretty dog-eared by the time we got to it. On the other hand, Dr Halstead's plea for "scholarship" seems equally pathetic. He must realise that not many of us would come up with an evolutionary theory from a perusal of a Cuverian arrangement of specimens, while I cannot imagine what diabolical deeds he envisages the "more inarticulate sections of the community" being incensed to perpetrate as a result of exposure to the Natural History Museum's "social engineering".

However, I feel much more sympathy for Halstead's main worry of a "serious breakdown of communication between the museum's Department of Public Services and the scientific staff . . .". It's all very well to cite the dismay of the museum's "world leading ichthyologist" and "dinosaur expert" at the loss of the exhibits of their respective groups, but they didn't reach their present eminent positions on the strength of their efforts in the exhibition sector. For like the rest of the scientific staff they know that their professional advancement has depended entirely on publication; time devoted to exhibition work has therefore tended to be spent grudgingly and, at least in the past, was often restricted to topping up a dried out bottle or replacing a lost label. Under such circumstances it is little wonder that a them-and-us situation developed between the public and private sections of the staff, with the scientists taking little or no interest in the exhibits for which they were nominally responsible. The scientists can't expect to have their cake and eat it, for if they want a real say in the fate of the exhibitions they must be prepared to be involved. In return, the museum establishment must demonstrate the value they attach to such involvement in terms of career structure and promotions. The Department of Public Services seems to have evolved to fill a gap resulting from the mutual distrust felt on precisely this point.

If there are now two quite distinct, but genuinely held, opinions of the type of exhibitions which the museum should mount in the future, it is surely not too late to exploit the best features of both views. The projected building programme within the museum's existing walls is supposed to provide four times the existing exhibition space over the next three decades. Such an expanse must be capable of housing the traditional scholarly exhibits favoured by Halstead, the exciting new conceptual approach advocated by Miles, and the fun exhibits which the dominant less than 11-year-olds can stare at in wonder and awe and remember in the recesses of their mind for the rest of their lives.

Yours faithfully,

A. L. RICE

Cranleigh, Surrey, UK

New policies are correct

SIR,—There seem to be three separate strands in the recent controversy on the new display policy of the British Museum (Natural History): whether displays should be for scholars or for the general public, whether they should be organised by research scientists or by exhibition staff, and whether the former balance of displays should be retained. In each case, I believe that the new policies of the BM(NH) are basically correct.

Until recently, the majority of the displays were little more than a collection of labelled specimens. There was little attempt to point out any features of particular interest, or to help the visitor to appreciate their significance to current scientific theory. Even those displays directed towards university students were suited to a level of knowledge that was not usually attained before the second year of a zoology degree. Similarly, those exhibits that were directed at the reasonably well-informed adult were accompanied by complex and lengthy labels, their comprehension requiring a considerable degree of dedication and time. Finally, many of the exhibits were so old-fashioned

as to merit illustration in any book of the history of museum display technique. It can be no surprise that the majority of the visitors to the museum had, as Dr Miles stated "learnt practically nothing". It is surely appropriate that the museum should attempt to inform, as well as interest, the throng of children and of scientifically uneducated adults who nowadays pack its halls.

Should we really expect that the prime responsibility for the new exhibition approach should rest with the scientific staff, already committed to programmes of research and curation? It is no secret that, in any museum, many research workers have little interest in the problems of the design of displays, even if they are of completely conservative and scholarly type. It is even more unlikely that many of them would wish to learn and to apply the theory and techniques of modern displays directed at those with no scientific qualifications. (This is not to say that the scientific staff should have no role in this, and many of those at the BM(NH) certainly seem to feel that they have not had sufficient opportunity to comment on the new policies, or to ensure that the scientific theories illustrated are appropriate and not merely those currently in vogue).

The balance of exhibits is a complex matter. I feel that it is only appropriate that the museum should attempt to let visitors understand themselves as human examples of the world of nature, and to educate them in the basic processes of ecology. (Not even the museum display staff would pretend that every exhibit in the Hall of Human Biology is an unqualified success, but even the BM(NH) must surely be allowed to make the occasional mistake). Furthermore, there can be no doubt that invertebrates and plants have in the past received a totally inadequate proportion of exhibition space. It was therefore inevitable, and only right, that the new exhibits should be at the expense of the great range of extensive (though mainly old-fashioned) displays of living and fossil vertebrates. It is natural for vertebrate zoologists to regret this, but we should surely be reassured by Dr Miles' statement that the new exhibits "will contain the vast majority of the material now on show in the museum".

Yours faithfully,

BARRY COX

King's College, London

Links with school

SIR,—As a teacher of biology, I would like to offer my views about the recent article entitled 'Whither the Natural History Museum?' by Dr B. Halstead.

My earliest recollections of the museum are those of my childhood, when I was taken on the usual tour of the museums on a trip to London. I remember, most of all, the dinosaurs but little else except endless rooms of glass cabinets full of bones and a lot of unpronounceable words.

Today, thankfully, the presentation of material in the museum is progressing and the two most prominent examples of this are the new exhibitions of human biology and ecology. These offer a new and very important dimension to the role of the museum—namely a vital resource which links important aspects of natural science with what is being taught at school.

I have been lucky enough to use the exhibition of human biology as part of my exam course in the fourth and fifth years. The pupils concerned were taught how to use the exhibition and, as a result, were able to benefit from it in a way they could not by just relying on textbooks at school. I look forward to using the ecology exhibition in the same way with my second year groups next summer. The publications accompanying the two exhibitions also allow for follow-up work at school.

From Dr Miles' figures regarding visitors to the museum, it would appear that these exhibitions will benefit an age group whose presence is sadly missing from the museum.

I would like to say in conclusion that the museum should continue to present its rare and fascinating fossil exhibits, as it has done for years, thus fulfilling its role to the scholars; but also to continue with its new fresh approach and present more exhibitions which are of more use to schools and the less scholarly members of the general public.

Yours faithfully,

D. A. WEALE

St John's School, Epping, Essex, UK

Teachers need a say

SIR,—I am a teacher in the fourth year of a large multi-cultural East London primary school. I am prompted to write in response to the articles by Dr Miles and Dr Halstead on the future of the British Museum (Natural History).

Dr Miles says that to some extent the Natural History Museum is for children under the age of eleven. He points to the absence of secondary school pupils and university students from the public galleries. Without doubt the primary school children are impressed by the large dinosaurs and blue whales. They take a passing interest in one or two other galleries or individual exhibits but generally scurry through the remainder of the museum if given the opportunity. To the children many of the current displaps seem unrelated to each other or to a wider theme. I am not surprised that the majority of younger visitors are not motivated to return for many years.

On the other hand the recent human biology exhibition is stimulating and dynamic and makes demands of the children. In my experience they take an active interest in the displays and are keen to seek answers to questions. Furthermore the exhibition is coherent and so more meaningful to them. Although many of the concepts explained are beyond the grasp of the average eleven year old, I anticipate that a higher proportion are likely to revisit and develop their understanding than would have done had the exhibition remained in the traditional layout. If other recommendations in the paper 'A Proposal for a New Approach to the Visiting Public' are implemented I feel that many teachers and pupils will benefit even more.

It is disquieting to hear of the rift between the Department of Public Services and scientific staff. Dr Halstead himself, however, refers to a false dichotomy, by discussing the former dinosaur gallery and its multi-level appeal. Surely the 'new style' exhibition can also be both stimulating and scholarly given the goodwill and collaboration of scientists and exhibition staff. After all Dr Miles does say that the exhibits to be displayed will contain the vast majority of the material on show.

Dr Halstead calls for the fullest possible public discussion. Hopefully teachers at all levels will be given an opportunity to participate in this discussion which at the moment seems to be conducted at the level of the academic.

Yours faithfully,

L. J. STEVENS

Wanstead Park, London, UK

Arborescent animals and other colonoids

SIR,—Comparative biology is sometimes hindered by terms which have become misleading with the advance of knowledge. 'Individual' and 'colony' are such terms when they are used to refer to a single polyp and the entire organism of a reef coral, for instance. The organism responds as a unit in growth and physiology, and it is also the unit of individual selection. I therefore propose two new general terms to replace 'individual' and 'colony' in such cases. Specific terms are already available in some instances, but they vary from taxon to taxon.

Individuoids are parts of an organism which have the general structure of whole free-living individuals but which connect with each other to form a coloniod. A colonoid usually has the same genes throughout and functions as a single individual.

The concept of colonoid grades into that of colony or association, which may be restricted to a group of proximate but physiologically separate individuals, usually of one clone. Intermediate cases might be grass swards or diatom chains. The concept of a colonoid organism also grades into that of individual *sensu stricto*, in which case individuoids may resemble organs. Examples are rooting branches of mangroves, bryozoan avicularlae, or cyanophyte heterocysts. Slime mold colonoids are colonoids for only part of the life cycle and may form (sometimes necessarily so) from individuoids of different genotypes. Such fuzzy boundaries of concepts are necessary in the real world.

A diverse array of animals grow in a treelike form. Some but not all have apical meristems. Examples of arborescent growth occur among sponges, hydroids, alcyonarians, corals, bryozoans (including entoprocts[1]), graptolites, pterobranchs, and probably elsewhere. Comparative studies of their growth and that of branched plants and protozoans, from both mechanistic and adaptive viewpoints, have always proved illuminating in my seminar course on the evolution of development and would undoubtedly be more illuminating if pursued more fully.

The study of plants and arborescent animals can thus be mutually beneficial, but such study presupposes that functionally corresponding units are identified beforehand. Studies of other phenomena, such as adaptive strategies, also need clear recognition of what units are appropriate. This has not always been done, perhaps because of the pervasive shadow of a purely morphologically based terminology, and the result is then biological nonsense.

Yours faithfully,

LEIGH VAN VALEN

Biology Department, University of Chicago, USA

[1]Neilson, C. *Ophelia*, 9, 209-341; 1971.

news and views

Whaling: past, present and future

from Robert M. May

FEW, if any, conservation movements have received as sustained and widespread support as those aimed at 'saving the whales'. Although most of us never have and never will see a great whale outside a museum, we are dismayed by the excessive exploitation these all-time largest of animals have suffered.

People who are vaguely aware of the sorry history whereby stock after stock of whales has been harvested well below the point of maximum sustainable yield (MSY) may be surprised to learn how things stand today. Of the twelve species of whale that have been hunted over the centuries, seven are completely protected in all oceans, three in most ocean areas, and the remaining two in some areas, under the current policies of the International Whaling Commission (IWC). There remain, of course, many problems, not the least being whaling activities under the flags of nations that are not IWC members, and the fact that even for whaling by IWC members the enforcement of agreed regulations leaves a lot to be desired.

The New Management Procedures of the IWC were adopted at its 26th meeting in 1974, and first put into effect in the southern summer of 1975–76. This procedure divides whale stocks in each of the ocean areas defined by the IWC into one of three categories, according to whether the populations are thought to be numerically bigger or smaller than, or near to, the size required to give MSY (for a more full account of the notion of MSY, see *News and Views* **263**, 91; 1976). This classification is revised annually, and more or less automatic rules then applied; if the stock is judged to be below the MSY level, a moratorium is declared ('protection stocks'); if it is near the MSY level, annual quotas are set with the intent of keeping the stock around this level ('sustained management stock'); if the estimated stock size is significantly above the MSY level, exploitation is permitted under quotas aimed to produce a controlled reduction toward the MSY level ('initial management stock'). Application of these rules in 1978 (and in the summer 1977–78 in the southern oceans) produced the numerical quotas shown in the maps. In addition to the wholly protected blue, grey, humpback, and four species of right whales, it will be seen that fin, sei and Bryde's whale are protected in most areas. Sperm whales have significant sexual dimorphism and a more elaborate social structure than the baleen whales, and for them separate quotas are set for males and females (see legend); the models used to obtain these numbers are currently being scrutinised and refined by the IWC.

Before considering the strengths and weaknesses of this management procedure, it is interesting to look at some of the historical background and at the tensions between economic interests and sustainable yield criteria.

History of whaling regulations

The whales that have been hunted include 11 species of baleen whales, which are filter-feeders with baleen plates ('whalebone') in place of teeth: (see box). The largest species of toothed whale, the sperm whale, *Physeter catodon*, is also hunted, and provides most of the present world catch by numbers and by weight (though not by value).

Formerly, these animals provided oil for lamps and lubricants, and the baleen whales also yielded whalebone.

Bowhead (Greenland right)	*Balaena mysticetus*
North Atlantic right	*Eubalaena glacials*
North Pacific right	*Eubalaena seiboldii*
Southern right	*Eubalaena australis*
Grey	*Eschrichtius robustus*
Blue	*Balaenoptera musculus*
Fin	*Balaenoptera physalus*
Sei	*Balaenoptera borealis*
Bryde's	*Balaenoptera edeni*
Minke	*Balaenoptera acutorostrata*
Humpback	*Megaptera novaeangliae*

Baleen whales remain a source of oil for the manufacture of margarine, soap, and glycerine, of meat which is eaten directly by humans and pets, or used in soups and flavouring and of meat and bone meals for fertilisers and animal foods. Sperm whale oil and spermaceti provide industrial scouring and hardening agents, lubricants and cosmetics. As emphasised by many people, "all these products could be replaced by synthetic derivatives from vegetable and mineral sources, but it is still economically attractive to produce them from whales" (Gambell & Brown *IUCN Bull. Ser. II* **21**, 1; 1971).

The North Atlantic right whale was hunted by the Basques as early as the twelfth century. By the beginning of the twentieth century, the more easily catchable species—grey, three species of right, bowhead (the latter four species often lumped together as right whales)—were reduced to very low levels of abundance. These are all slow swimmers that usually float when dead, and they were pursued in open boats under oars or sail; real Moby Dick stuff. Whaling for the swifter swimming rorquals—blue, fin, sei, Bryde's and very recently minke—began in the mid-nineteenth century with the invention of the explosive harpoon, harpoon cannon, and the use of steam-powered ships. It spread to the Antarctic in 1904, starting at South Georgia. The advent of the factory ship in the mid-1920s freed whalers from land stations, and began the modern era of Antarctic pelagic whaling. By 1935, fears that Antarctic stock would be depleted as Northern ones had earlier been culminated in the formation of an International Convention for the Regulation of Whaling, which banned all harvesting of right whales or of any female whales with calves.

World War II gave the whales a respite. In 1946, the pre-war threads were picked up by a new International Convention which established the IWC, charged to "provide for the conservation, development, and optimum utilisation of the whale resources, [taking]

Robert M. May is Class of 1877 Professor of Zoology at Princeton University.

0028–0836/78/0276—0319$01.00

© Macmillan Journals Ltd 1978

into consideration the interests of the consumers of whale products and the whaling industry". The grey whale was added to the protected list in 1947.

The Commission is advised by a Scientific Committee, which from early days has issued warnings and pleas that stocks were being exploited too heavily. The present analytical models and quantitative methods of population assessment derive mainly from the work of the Committee of Three Scientists (Allen, Chapman and Holt), appointed by the IWC in 1960. The reports of this committee in the early 1960s made it unanswerably clear that stocks of blue, fin, and humpback whales were being seriously overfished, and in 1963 they recommended total protection of blue and humpback, and a reduction of the Antarctic catch of fin to 5,000 whales or less. The IWC responded by partly protecting the blue and humpback, and reducing the Antarctic catch limit from 15,000 to 10,000 'Blue Whale Units' (this was an atrocious unit, whereby 1 BWU=1 blue, or 2 fin, or 2.5 humpback, or 6 sei). This pattern of the IWC reacting to the advice of its own Scientific Committee by doing too little, too late, persisted. Nonetheless, the 1960s and early 1970s saw some advances as the seeds of the New Management Procedures were sown: the Blue Whale Unit was abolished; two species were added to the list of those completely protected (the blue in the North Atlantic in 1960, the North Pacific in 1966 and world-wide in 1967, and the humpback in the North Atlantic in 1955, southern hemisphere in 1964, and world-wide in 1966); quotas for the Antarctic fin and sei were eventually set under the constraint that they be no more than the estimated sustainable yield.

Outside the IWC, pressures were building up. In 1966, the 1958 Geneva Convention on the Conservation of the Living Resources of the High Seas came into force, adjuring "measures rendering possible the optimum sustainable yield . . . so as to secure a maximum supply of food and other marine products". The UN Conference on the Human Environment, held in Stockholm in 1972, recommended a 10 year moratorium on all commercial whaling, and this recommendation was subsequently adopted by the UN General Assembly and by the General Assembly of the International Union for the Conservation of Nature (IUCN).

The IWC considered this recommendation at its meeting in 1974, but settled for adopting the New Management Procedures described above. The technical details were hammered out in 1975, and led to the following definitions. A protection stock is one which is more than 10% of MSY stock level (MSYL) below the MSY level itself. A sustained management stock is one that is not more than 10% of MSYL below, nor 20% of MSYL above, the MSY level. Finally, an initial management stock is one which is estimated to lie more than 20% of MSYL above MSYL. Thus, for example, if the stock obeyed a logistic recruitment curve, and if the virgin stock density were defined to be 1, then a stock density of less than 0.45 would be protected, between 0.45 and 0.60 would be sustained management, and greater than 0.60 would be initial management. The rules also embody some caution with respect to possible errors in the estimates of MSY levels; no more than 90% of the estimated MSY can be taken, regardless of the stock class. Furthermore, for initial management stock, not more than 5% of the estimated initial exploitable stock is to be taken in any one year, unless there is 'positive evidence' that higher levels will not lead to over-exploitation. Nor may significant exploitation begin until a satisfactory estimate of stock size has been obtained. (For more full accounts, see Brown *Polar Record* **18**, 85; 1976; **19**, 59; 1978; *FAO Fisheries Report* No. 194 (the Report of the Advisory Committee on Marine Resources Research, chaired by Laws); 1977.)

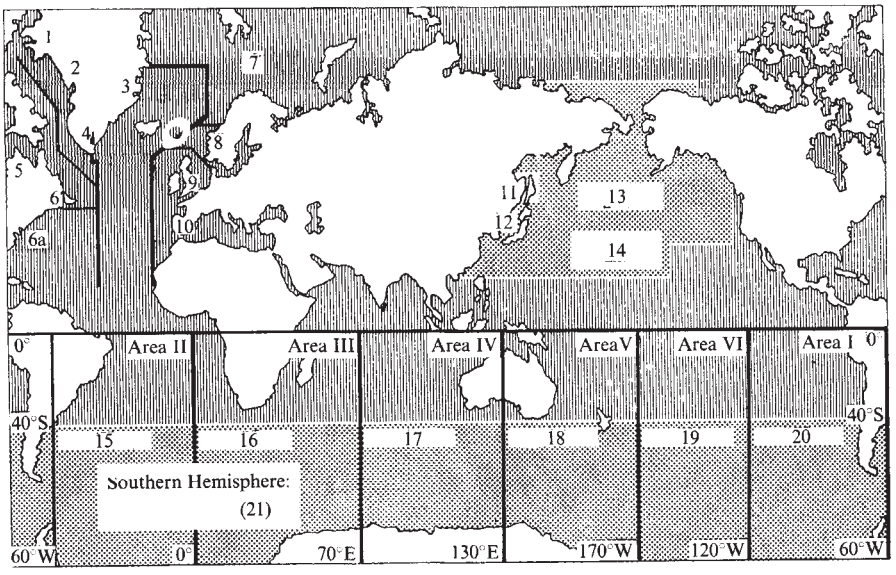


Fig. 1 Quotas for baleen whale species in the various ocean areas designated by the IWC (1977/8 and 1978 seasons). PS stands for protection stock, SMS for sustained management stock, and IMS for initial management stock, as discussed more fully in the text. (1) West Greenland: minke (SMS) 397*, fin (SMS) 4*. (2) Iceland-Denmark Strait: sei (SMS) 84. (3) East Greenland-Iceland: fin (SMS) 304. Not more than 1,524, 1977-1982. (4) East Greenland-Iceland-Jan Mayen: minke (SMS) 320. (5) Canadian East Coast: minke (SMS) 48. (6) Newfoundland-Labrador: fin (IMS) 90. (6a) Nova Scotia: fin (PS) 0; sei (PS) 0. (7) North Norway: fin (SMS) 61*. (8) West Norway-Faroe Islands: fin (PS) 0. (9) Svalbard-Norway-British Isles: minke (SMS) 1,790. (10) Spain-Portugal-British Isles: fin (SMS)*. (11) Okhotsk Sea, West Pacific: minke (SMS) 400. (12) Sea of Japan: minke (SMS) –. (13) North Pacific: fin, sei (PS) 0, Bryde's (IMS) 524. (14) Remainder: minke (IMS) 0 pending a satisfactory estimate of stock size. (15) Sei (PS) 0, minke, (–) 1,150. (16) Sei (PS) 0, minke (–) 1,826. (17) Sei (SMS) 460, minke (–) 963. (18) Sei (PS) 0, minke (–) 930. (19) Sei (PS) 0, minke (–) 688. (20) Sei (SMS) 388, minke (–) 704. (21) fin (PS) 0; Bryde's (IMS) 0 pending a satisfactory estimate of stock size; sei, catch not to exceed a total of 771; minke, catch not to exceed a total of 5,690.

MSY *versus* economic considerations

Reviewing the above history, Gambell and Brown (*op. cit.*) note that "despite warnings from the scientists that the catches were considerably larger than the stocks could withstand in the long term, the whaling nations allowed their industries to take almost as many whales as they could, arguing that they could not otherwise survive economically. They leaned on the fact that the scientific evidence was not complete, precise and incontrovertible". The basic reason for this behaviour is that, although the discussions formally centre around sustainable yield (with present and future catches in principle regarded equally), the industry is more interested in economic returns. This tension between MSY principles and economic actualities takes many forms, some of which are widely appreciated and others not; the general issues have been set out incisively by Clark (*Mathematical Bioeconomics*, Wiley, 1976).

The first problem lies in establishing some agency that is effectively the sole owner of the resource. Failing this, an open-access fishery suffers a 'tragedy of the commons' (Hardin *Science* **162**,

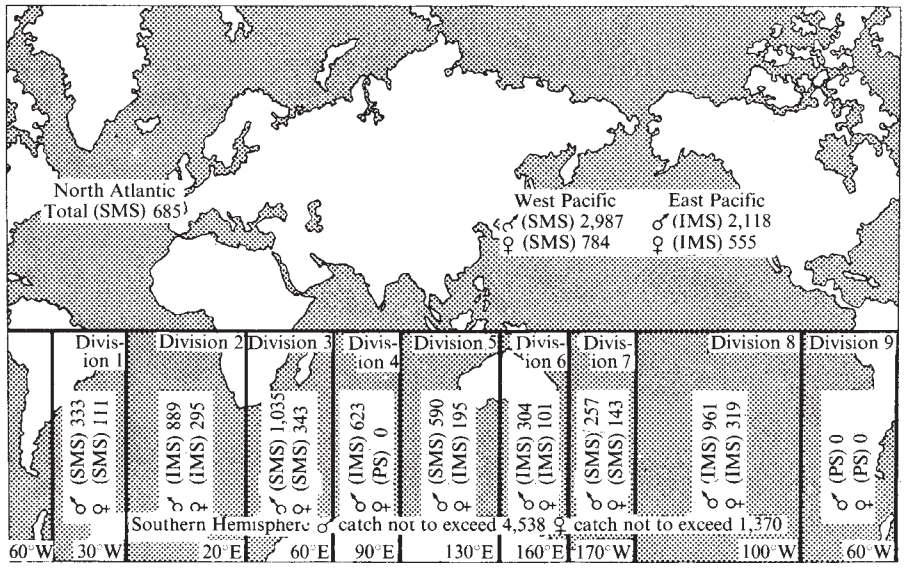


Fig. 2 IWC quotas for sperm whales (1977/78 and 1978 seasons); there are separate quotas for males and females.

1243; 1968), in which the 'economic rent' is dissipated by exploitation down to the point where revenues are balanced by harvesting costs. The establishment of the IWC, which aims to act as such a sole owner, is thus a non-trivial accomplishment for which its members deserve credit. It is in the nature of the situation that there remains a tendency for individual ships (and even—whisper it—nations) to cheat.

Moreover, as stressed by Clark (see also *News and Views op. cit.*), whalers are interested not in maximising sustainable biological yields, but in maximising the present value (PV) of net economic return, discounting the future. If the economic discount rate is small compared with the biological growth rate of the exploited stock, these considerations are relatively unimportant, and MSY and PV criteria lie close. But if the economic discount rate is greater than, or of the order of, the population's intrinsic growth rate (as arguably is the case for whales, where growth rates lie around 5 to 10% per annum), the PV criterion will tend to lead to stocks being depleted below the MSY level. Indeed in the limit when the discount rate becomes infinite (which in practice means when it significantly exceeds the biological growth rate) exploitation by a sole owner is mathematically identical with the open-access case. The reason is intuitively clear; with relatively high discount rates there is, just as in open-access exploitation, no motive to heed the future. These remarks are subject to many qualifications, among them that prices, harvesting costs, and inflation-corrected discount rates are all liable to change over time (Clark, *op. cit.;* Gulland, *Nature* **263,** 722; 1976; Clark & Munro *J. Env. Econ. Mgmt.* **5,** 198; 1978). All this can make for great complications in the analysis, but I think the essentials are as above.

These economic considerations have some subtle corollaries which are not so widely acknowledged. Clark and Munro (*J. Env. Econ. Mgmt.* **2,** 92; 1975), for example, have shown that the optimum trajectory for capital investment can, under some circumstances, lead to an initial rapid phase of heavy exploitation, possibly even depressing stock below the level that maximises PV (much less SY), followed by a slower phase in which the stock levels recover to and are maintained around the PV level. Very crudely speaking, these two phases correspond to fast capitalisation followed by relatively slow amortisation of the equipment.

Even within an agreed MSY framework, there is room for squabbles motivated by economics. One such dispute at present concerns whether the quotas for sperm whales should be set for MSY by numbers or by weight. It happens that for sperm whales a significant fraction of the life-long growth in size takes place during the exploited adult phase of the life cycle. The result is that, once the equilibrium age structure is attained under a given harvesting regime, MSY by weight gives a total catch greater in biomass (and thus in products) and lower in numbers (therefore requiring less effort and lower harvesting costs) than is obtained by MSY based on numbers (Holt, FAO Document ACMRR/MM/SC/99; 1976). Unambiguously, the long term advantage lies with MSY by weight. The snag is that several years must elapse before sperm whale populations change from the existing age structure, produced by harvesting for MSY by numbers, to that appropriate to the MSY weight criterion. If the future is heavily discounted, one will always argue for the numbers criterion, because this will allow more whales to be caught, and, in the immediate future, their average weights will be much the same under either quota system. In this way can the future be squandered even in an MSY context.

One bleak consolation lurking in the above discussion is the possibility that whaling, once it is effectively controlled at MSY (as opposed to maximum PV) levels, may cease to be economically attractive. This issue, however, requires taking a longer view of prices and costs than I am prepared to do.

Prospects for the whales

The present biomass of large whales is estimated to be about 23 million tons, which is about one-third the value around 1900 (FAO *op. cit.*). Some species have, of course, suffered very heavily; for example the populations of blue and humpback are down to a few percent of their pristine numbers. However, no species has been extinguished by whaling, although the grey whale is extinct in the North Atlantic.

Ironically, the one species that could conceivably be in danger of extinction is the bowhead. Although this species has long had protection status, up to 1978 the IWC rules permitted hunting by aboriginal peoples. Recent increases in the prosperity of societies of Alaskan Eskimos have resulted in their hunting with more boats, which seem on the average to be less ably manned. Not only have the number of bowhead caught increased markedly over the past decade (from an average around 9 per annum in 1912-1967 to around 27 over the past decade: Clapman, *Marine Mammal Information,* Dec. 1977), but the number struck and lost has greatly increased. It is likely, although in dispute, that the number of animals thus killed exceeds the number actually harvested. Conservative estimates suggest the present levels of exploitation exceed the rate of net reproduction. The sentimental argument that these people should be allowed to continue their depredations, lest their traditional culture be imperilled, seems to me to miss the point that the current problems arise largely because the traditional culture is already collapsing.

The only stock that has demonstrably recovered under protection is the grey whale in the eastern North Pacific (the California grey whale). Some humpback and some right whale populations show indications of increases in recent years. It is still too early to detect with assurance whether blue whale stocks have been increasing since

they were protected. The uncertainties here derive from the difficulties in getting enough survey data to measure significant changes in abundance.

The New Management Procedures have the virtues of recognising individual species and geographical stock units, each subject to its own management regime, and of classifying stock with the explicit aim of long term conservation. The allowance for a margin of error in the estimation of MSY level is another step in the right direction; this precaution helps offset uncertainties in the determination of population parameters, and gives some buffering against the unpredictable environmental fluctuations to which all natural populations are subject to one degree or another. The procedures thus go some way toward meeting the broad principles for conservation and management articulated in the so-called Airlie proposals, which emerged from a consultation sponsored by the World Wildlife Fund, the Ecological Society of America, the IUCN and others (see Holt & Talbot *Wildlife Monogr.* No. 59; 1978).

The New Management Procedures could be improved in at least three ways. First, they could explicitly acknowledge the long time lags inherent in the dynamics of whale populations, and the problems these cause. Second, they should aim to maximise the net sustainable yield of whale products rather than the number of whales caught; that is, they should take full account of whaling costs, and should maximise the catch by weight not by numbers. Finally, and most difficult, it needs to be recognised that whale stocks are not single species, but are constituents of a multispecies ecosystem. A management procedure that (as required by the US Marine Mammal Protection Act of 1972) takes due account of the 'health and stability of the ecosystem' will need to grapple both with technical issues of multispecies MSY criteria, and with broad ecological questions of the possible existence of alternative stable states.

Immunoglobin genes: more facts to guide speculation

from W. A. Dunnick

BIOLOGISTS have proposed all sorts of genetic gymnastics for immunoglobulin genes. They have suggested that immunoglobulin genes hop around chromosomes, and recombine or mutate at terrifying rates. The genes have been amplified, branched, interrupted by episomes, and somatically diversified, but never understood. Immunoglobulin genes attract this bizarre speculation because of the protein's enigmatic structure.

Both the light and heavy chains found in immunoglobulin proteins have structurally simple carboxtermini. For example, the carboxy halves (C-regions) of about 95% of mouse light chains have an identical 'kappa' sequence. The remaining 5% are divided between two related 'lambda' sequences. On the other hand, the amino terminus (V-region) is very heterogeneous in sequence. Individual mice may be able to express 10,000 different V-regions. The variability is not random, but is concentrated in three 'hypervariable' regions No other group of polypeptide sequences implies such an extensive, yet nonrandom, divergence of genes. Most speculation concerning immunoglobuln genes centres on two problems: (i) the bringing together of the very different V and C sequences to form a single polypeptide, and (ii) the generation of V-region diversity.

V–C joining

If each of the 10,000 V-genes were associated with a C-gene, the first problem would be solved. However genetic and gene counting experiments have shown not thousands, but only a few C-genes. Thus, this simple model must be excluded.

A second simple model would be the association of a few C-genes with a few V-genes. Mutations in the V-genes during somatic development could give rise to 10,000 different V sequences (in 10,000 different cells) associated with a single C sequence. However, there seem to be more than a few germline V-genes. V-regions are grouped according to shared amino acid sequences; members of a single subgroup share three or more residues not found in other subgroups. To avoid the unpalatable assumption of parallel mutation, each subgroup is thought to be coded by at least one germline gene. Such an analysis implies that the few kappa C-genes must accommodate 50 or more kappa V-genes.

Dreyer and Bennet (*Proc. natn. Acad. Sci. U.S.A.* **54**, 864; 1965) proposed that V and C-genes are brought together by a translocation event, cells expressing different V-regions rearranging different V-genes. Hozumi and Tonegawa (*Proc. natn. Acad. Sci. U.S.A.* **73**, 3628; 1976) demonstrated such a rearrangement for a mouse kappa chain (see *News and Views* **264**, 699; 1976), and rearrangement of V and C-genes has also been shown for a second kappa chain (Rabbitts & Forster *Cell* **13**, 319; 1978) and for a lambda chain (Tonegawa *et al. Cold Spring Harbor Symp. quant. Biol.* **41** 877; 1976).

These observations suggested that translocation results in the contiguous V–C sequence found in mRNA (Milstein *et al. Nature* **252**, 354; 1974). It was soon demonstrated that this is not the case. Nucleic acid hybrids between kappa cDNA (Rabbitts & Forster, *op. cit.*) or mRNA (Matthyssens & Tonegawa *Nature* **273**, 763; 1978) and their respective myeloma genes have looped-out, single-stranded regions that are susceptible to S1 nuclease. The size of the products of S1 digestion suggests that the discontinuity in the gene is near the V–C junction. Also, electron microscopy of R-loops formed between lambda mRNA and a cloned myeloma V+C gene reveal a 1,200 base pair intervening sequence between the V and C coding sequences (Brack & Tonegawa *Proc. natn. Acad. Sci. U.S.A.* **74**, 5652; 1977).

The complete structure of the mouse lambda germline gene (Bernard, Hozumi & Tonegawa *Cell*, in the press) has now been determined. It consists of 45 bp coding for the lambda signal peptide, 306 bp of variable region, more than 4,500 bp of undefined DNA, 39 bp of variable region sequence, approximately 1,200 bp of intervening sequence, 312 bp of constant region sequence. (The preceding implies that the V-gene is 'upstream' relative to the C-gene; this is not known to be the case.) In cells producing lambda light chains, a translocation event fuses the two V segments, eliminating the large undefined segment of DNA from the active gene. The small segment of 'variable' region associated with the C-gene (albeit 1,200 bp away) is a surprise. Bernard *et al.* have termed the segment J, emphasising its role in V–C joining. J. Seidman and colleagues (NIH) reported a similar structure for a kappa gene at the Seventh Workshop on Homogenous Antibodies*.

We are left with the V and C-genes over 1,000 bp apart in myeloma DNA,

W. A. Dunnick is a Jane Coffin Childs Memorial Fund Postdoctoral Fellow at the MRC Laboratory of Molecular Biology, Cambridge.

*Held at the National Institutes of Health, 23–25 October.

0028–0836/78/0276—0322$01.00

© Macmillan Journals Ltd 1978

but contiguous in mRNA. Three lines of evidence indicate that nuclear RNA precursors (Gilmore-Hebert & Wall *Proc. natn. Acad. Sci. U.S.A.* **75**, 342; 1978; see also Schibler, Marcu & Perry *Cell* in the press) contain these intervening sequences. First, hybrids between cDNA and the precursor RNA are not perfect matches, but have single-stranded regions near the V–C junction (Rabbitts *Nature* **275**, 291; 1978). Second, the precursor and mRNA apparently have the same nucleotide sequence in the C-region. On the 5′ side of the C-region, sequences in the precursor differ from those in the mRNA (Rabbitts *op. cit.*). Third, ultraviolet inactivation experiments suggest that V and C coding sequences are not adjacent in nuclear RNA (R. Wall & M. Gilmore-Hebert, University of California, Los Angeles, reported at Homogenous Antibodies Workshop). The intervening sequences in nuclear RNA must be removed by some RNA processing mechanism.

Single cells produce only one sort of light chain, but seem to be able to couple one heavy chain V-region with two or more heavy chain C-regions (for example, mu and gamma C-regions). Several authors have suggested that the heavy chain V-region may be cotranscribed with several C-regions, and that RNA processing would determine which heavy chain V–C combination(s) is to be expressed. This elegantly simple idea will have to be revised; hybrid cells producing two different classes of heavy chain do not exchange V and C-regions (Milstein *et al. Cold Spring Harbor Symp. quant. Biol.* **41**, 793; 1977; Köhler & Schulman *Nature* **274**, 917; 1978). For example, hypothetical 'mu' processing enzymes do not bring together the mu constant region and a variable region associated with a gamma constant region, even though both genes are transcribed in the same hybrid cell. In addition, R. Perry, U. Schibler and K. Marcu (Institute for Cancer Research, Philadelphia) reported at the Homogenous Antibodies Workshop that both gamma and alpha heavy chain mRNAs are transcribed as precursors at least 8 to 11 kb long. Nevertheless, neither mu nor alpha transcripts are detected in the nucleus of the gamma producer, and neither mu nor gamma transcripts are found in the nucleus of the alpha producer. Thus, the mu-gamma-alpha switch cannot be explained by a simple RNA processing model (for an alternative theory, see Honjo & Kataoka, *Proc. natn. Acad. Sci. U.S.A.* **75**, 2140; 1978).

V-region diversity

The extraordinary results discussed above have solved, in a general way, many of the problems of light chain V–C joining. In relation to V-region diversity, many immunologists believe that some number of germline V-genes (say 100 to 1,000) diversify somatically to yield a larger number of total genes (10,000). However, this is not to say that all immunologists endorse such a statement; the gap between belief and proof has been great. Kinetics of hybridisation experiments have suggested a limited number of germline V-genes, but the interpretation of such data is very difficult (see Smith *Cold Spring Harbor Symp. quant. Biol.* **41**, 863; 1977; Rabbitts & Milstein *Cont. Topics Molec. Immun.* **6**, 117; 1977).

The extensive study of one kappa group (V21) has been more informative than investigations of the widely diverse kappa chains as a whole, or studies of kappa groups in which only a few amino acid sequences are known. The V21 group may be divided into at least six subgroups on the basis of related sequence patterns (McKean *et al. Proc. natn. Acad Sci. U.S.A.* **75**, 3913; 1978; Weigert *et al. Nature*, in the press). The subgroup structure is most clear if residues 99–110 (J-region) of the V-regions are ignored. The V21 subgroups seem to randomly associate with eight J-regions (Weigert *et al. op. cit.*) As Bernard *et al.* (*op. cit.*) proposed, V-region diversity may increase through recombination of V and J-regions or through imperfect V-J joining. Using a hybridisation technique (saturation) which should be more accurate than kinetic measurements, Valbuenna *et al.* (*Nature*, in press) have suggested that the number of V21 germline genes is the same as the number of subgroups (four to six). This would strongly favour somatic variation in this group, since nearly 30 amino acid sequences have been determined. Impurities in the cDNA preparation complicate the interpretation of these results, and one eagerly awaits the results of such experiments with cloned cDNAs.

Seidman *et al.* (*Science* **202**, 11; 1978) have also suggested that mouse kappa V-genes are represented in the germline by groups of at least six closely related genes. Two-dimensional 'fingerprints' of mouse DNA reveal six to ten bands of hybridisation to each of two different cloned kappa probes. Even though the two probes identify the same C-region band, they have no V bands in common. The authors speculate that the total number of kappa germline V-genes is 200 to 1,000. Experiments with many light chain probes will be necessary to reach a more precise estimate.

Cloning and sequencing of two V-genes hybridising to the MOPC 149 kappa probe have demonstrated that the two genes vary from each other and from the MOPC 149 V-gene by about 7%. The two V-genes are also homologous over 2,500 bp of flanking sequence. Seidman *et al.* expand an earlier hypothesis of Edelman and Gally (*Proc. natn. Acad. Sci. U.S.A.* **57**, 353; 1967). V-region diversity is expanded by recombination, the recombination being facilitated by the homology between flanking regions of different V-genes. The hypothesis is especially persuasive, since other structural genes (β-globins) diverge more slowly, and have less flanking region homology.

In spite of the abundance of remarkable data, our concepts of immunoglobulin genes remain vague in several areas. Are any special mutational mechanisms required to generate V diversity? How are heavy chain genes rearranged and joined? What is the molecular basis of immunological memory and tolerance? And finally, cellular immunologists spend more time with T cells than with immunoglobulin-producing cells these days. How do T-cell receptor genes relate to Ig genes. □

Cosmic spherules

from D. W. Parkin

MOST scientists will not have heard of cosmic spherules, yet Murray and Renard found and named them long ago (*"Challenger" Reports* **4**, 1891). About a dozen of these extraterrestrial spherules, ~ 100 μm diameter, can be magnetically extracted from a cupful of deep oceanic red-clay, which slowly deposits at a few metres per 10^6 years. They are black and quite spherical. Two types occur: the 'stones' and the 'irons', although a few might be called 'stony-irons'. The stones are olivine $(Mg, Fe)_2SiO_4$ and magnetic Fe_3O_4; a little glass may be present but not as a noticeable phase. The irons are magnetite and wüstite FeO; often a globule of Fe/Ni alloy is present, eccentrically placed in the oxide shell. There are other rounded magnetite forms in the red-clay which are not extraterrestrial. These occur especially in the less than 30 μm-sized range. Probably these faceted and very shiny black forms are growths on the sea bed or may come from volcanoes. Care is needed not to confuse these small particles with cosmic spherules. Above 50 μm there is little chance of contaminating a collection. The fact that the ratio of stones to irons is the same in any surface sample of red-clay provides a strong argument for an extraterrestrial source for both types of spherules.

Most meteorite experts would accept the extraterrestrial origin; but they

D. W. Parkin is at the School of Physics of the University of Bath.

© Macmillan Journals Ltd 1978

have tended to ignore cosmic spherules, believing that they are merely droplets sprayed from a meteorite as it ablates in the atmosphere. On the other hand, Opik (*Irish Astr. Jour.* **1**, 145; 1951; **4**, 84; 1956) proposed a more interesting origin: cosmic spherules are formed in the atmosphere by the melting of similarly sized fragments coming directly from the zodiacal cloud. Using his micrometeorite formula, he was able to predict size-distributions for cosmic spherules. Unfortunately, when the predicted and observed distributions are compared, the fit is not exact. Most zodiacal cloud particles are known to orbit the Sun with low inclination to the plane of the ecliptic, but it is not known how many of them depart from circular orbits. Therefore, there is some flexibility in the allowed velocity of impact with the Earth's upper atmosphere; also the shape of the impacting fragment enters into the theory. With contrived and somewhat artificial mixtures of velocities and/or shapes, a better fit can be achieved. Hunter and Parkin (*Proc. Roy. Soc.* **A255**, 382; 1960) largely avoided these complications by comparing the size distributions of stones with irons and they considered that Opik's zodiacal hypothesis was essentially correct, but the argument was not convincing. Meanwhile, to prove a zodiacal origin, interest was centred on making day by day airborne collections of cosmic spherules at remote sites; but these promising activities soon died away. The collections were always contaminated by industrial magnetic spherules. During the 1960s no positive progress favouring the zodiacal origin could be made; the rather uninteresting meteoritic ablation mode of formation was the only idea that prevailed.

Now, as a result of a few recent papers, interest in cosmic spherules is likely to be renewed. Brownlee *et al.* (*Interplanetary Dust and Zodiacal Light*, Eds. H. Elsässer and H. Fechtig, Springer-Verlag, 279, 1976) have painstakingly identified extraterrestrial dust by using aircraft in the stratosphere. Amongst these μm-sized crumbs, most of which are like carbonaceous meteoritic material, similar sized spherules have been found at about 10% frequency. It could be that both crumbs and spherules are coming directly from the zodiacal cloud. Parkin *et al.* (*Nature* **266**, 515; 1977), using scanning electron microscopy and X-ray diffraction on cosmic spherules, were persuaded that spherules are not meteoritic ablation droplets. From their surface texture, which shows only slight but varying atmospheric damage, it was claimed that cosmic spherules are already round bodies in space. With these ideas of non-molten (grazing) atmospheric flight, Opik's theory can be used to give a much more convincing explanation of the size-distributions, stones and irons being treated separately. If a spherule (on a plunging flight) was heated to melting temperatures, the size distribution analysis predicts that it would be destroyed; probably due to evolution of solar-wind gases accumulated during the spherule's residence in space. Very recently, Ganapathy *et al.* (*Science* **201**, 119; 1978) have analysed three stony spherules for trace elements, using neutron activation techniques. Not only do these analyses remove any lingering doubts about the extraterrestrial nature of stony spherules, but they break new ground in the chemical classification of cosmic spherules. This will become especially evident when more are analysed.

Of the eight elements analysed (Ir, Ru, Sc, Fe, Co, Ni, Cr, and Au) the amounts of Ir, Ru, Ni and Cr differ widely across the three spherules. Extra elements Zn, Sr, Os, Sb, Cs, Ag and Se are given for the third spherule. The authors also mention that S is highly depleted in these and other spherules. When the third spherule is compared with carbonaceous meteorite (C1) agreement is good, except for Au, Cs, Zn, and Ag which show moderate depletion. High depletion of very volatile Se and S is readily understood: the spherules have been molten at some time in their history. It is tempting (as the authors do) to compare cosmic spherules with C1 meteorites; and to attempt to derive each spherule from some primitive and homogeneous material, explaining the depletions and differences on the grounds of volatilisation on melting and leaching on residence in the sea water. But there is no compelling reason to do this for every spherule. It might turn out, when more spherules are analysed, that distinct families appear; with each family relating to the different meteorite types. It could even be that cosmic spherules are sparks from colliding meteorites or asteroids, the detailed chemical differences amongst them reflecting differences in the colliding bodies. Volatilisation of the moderately volatile elements or leaching of the soluble elements may not be as serious as Ganapathy *et al.* suspect—Cs is not highly depleted. In fact, we believe that most stony spherules in the surface sediments have not been severely attacked by sea water; the pores between the crystal grains being a natural freezing feature.

However, whatever the outcome of all this recent work, cosmic spherules are likely to become more widely talked about in future years. A pressing need is to find out where in the Solar System they are being created. This may be possible by examining deep-sea cores and using the fact that radiation from the Sun causes a drag on the orbiting spherule, the so-called Poynting–Robertson effect. Initial results from a red-clay core from the Pacific suggest the asteroidal belt. □

Muon spin rotation

from G. C. Stirling

ONE test of the public recognition of a new scientific technique is its emergence in the chemical sciences: optical spectroscopy, magnetic resonance spectroscopy and Mössbauer spectroscopy are common examples which have crossed the frontier from research curiosities in physics to routine tools in chemistry. The latest candidate is notable in representing spin-off from particle physics, a frequently-invoked, if trivial, justification for the vast funding accorded high energy physics research in recent years.

μSR—the acronym stands for Muon Spin Rotation, Relaxation, Resonance, Research or what have you, and refers to the analogy with NMR and ESR—has now come of age in this way after a relatively short adolescence, and at the recently-held first international meeting* on the subject much of the most promising new work related to purely chemical problems. Indeed, the use of muons in chemistry is striking not only for providing yet another probe of chemical systems, but for opening up an entirely new field of study—muonium chemistry—with important implications for related disciplines.

The μSR method in itself is conceptually simple, and derives from various serendipitous properties of the muon: it has a mass about 200 times that of the electron, a lifetime of 2.2 μs, and muon beams are inherently spin-polarised. Muons are rapidly stopped in bulk matter, and the spin direction is maintained during the thermalisation process. Subsequent interaction with local internal fields or with externally applied fields causes the muon spin to undergo a Larmor precession, and it is the precise behaviour of this precession which provides detailed information on the local muon environment. The precession is readily observed by the spontaneous muon

*First International Topical Meeting on Muon Spin Rotation, held at Rorschach, Switzerland, 4–7 September 1978.

© Macmillan Journals Ltd 1978

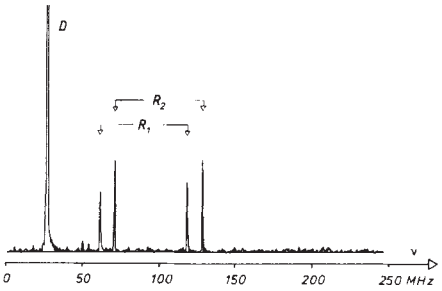


Fig. 1 Muon procession frequencies derived from the measured μSR time-histogram for 2-methyl-1,3-butadiene, with an external field of 1 kG. R_1 and R_2 originate from muonium-substituted radicals (after Roduner *et al., Chem. Phys. Lett.* **57**, 37; 1978).

decay, which leads preferentially to positron emission along the instantaneous muon spin direction, and appropriately located positron detectors give a direct observation of the time evolution of the muon polarisation.

Chemically the most interesting possibility is the formation and subsequent behaviour of the atomic species muonium ($Mu \equiv \mu^+ e^-$). Muonium has a similar size and ionisation potential to the hydrogen atom, and a mass some nine times less. It therefore would be expected to behave like a superlight hydrogen isotope, providing possibilities for observing dramatic H/Mu isotope effects in chemical reaction kinetic studies (compared with the relatively staid hydrogen-deuterium difference). Indeed, measurements of rate constants for the reaction $Mu+Br_2$ is the gas phase (Fleming *et al., J. Chem. Phys.* **64**, 1281; 1976), and Mu in aqueous solutions (Percival *et al., Chem. Phys. Lett.* **47**, 11; 1977; Jean *et al., Chem. Phys. Lett.* **57**, 293; 1978) have demonstrated the chemical similarity of H and Mu, although a detailed interpretation of observed isotope effects has been frustrated by the realisation that much of the previously accepted H atom data must now be reassessed. A new development reported at the Rorschach meeting concerns the direct observation of muonium-substituted radicals in solution. Although measurement of muonium rate constants obviously may imply the formation of such radicals, it is only recently that they have been observed directly in a μSR experiment (Roduner *et al., Chem. Phys. Lett.* **57**, 37; 1978). In muonium-substituted radicals the observed frequency spectrum is dependent on the coupling of the muon spin with those of nearby nuclei and unpaired electrons, leading to characteristic spectra. A typical example is shown in Fig. 1, corresponding to radicals $CH_2{=}C(CH_3)CHCH_2Mu$ and $CH_2MuC(CH_3)CH{=}CH_2$. Hyperfine coupling parameters are derived from the spectra, and the large isotope effects compared with the analogous hydrogenous radicals are related to equilibrium structure and vibrational and rotational motions of the radical species. The implications of this work for reaction rate theory and radiation chemistry are substantial, the importance of the μSR technique lying in its unique sensitivity—since only one muon may be in the sample at any one time, reactions are pseudo-first order, and the reacting muonium species can in principle be identified unambiguously from its depolarisation behaviour.

Why then, this apparently sudden activity in condensed matter applications of the muon? The answer undoubtedly lies in the availability of high quality muon beams at the new meson factories in North America, the USSR and Western Europe (though not, one notes, in the UK, although the proton synchrotron of the so-called spallation neutron source at the Rutherford Laboratory may yet provide a world ranking source). A striking feature of these facilities is the overall tendency towards 'applied' research, most markedly demonstrated by the Muon Spin Rotation technique itself. At the recent meeting for example, investigations using μSR included, in addition to purely chemical work, studies of the electronic structure of impurities in metals and magnetic materials, the behaviour of μ^+ and muonium in insulators and semiconductors, and the study of defects in radiation-damaged materials. As usual the biochemists have been quick to assess the potential of the technique in their own work. The proceedings are to be published in a volume of *Hyperfine Interactions* early in the new year, and this will doubtless serve to acquaint a larger audience with the possibilities of the technique for the first time. Already preliminary plans are in hand for a second international meeting, most probably to be held in Vancouver in 1980. □

G. C. Stirling is at the Rutherford Laboratory of the Science Research Council.

Superfluid ^{3}He in narrow cylinders

from Dieter Vollhardt

SUPERFLUID ^{3}He is a fascinating liquid —it has an unusually large number of internal degrees of freedom, it is inherently anisotropic and displays uncommon magnetic properties. There exist three superfluid phases: A, A_1 and a B phase which all differ in their spin-configurations. The A phase, in particular, can be characterised by two unit vectors: (1) the vector **l**, which describes the angular momentum of a ^{3}He–^{3}He Cooper pair and which is the anisotropy axis affecting, for example, flow motion of the liquid, and (2) the vector **d**, describing the spin part of the underlying wavefunction (if S is the total spin of a Cooper pair then $\mathbf{d}\cdot S$ is a constant of motion).

Both vectors are thus internally defined directions, which characterise the degrees of freedom of the superfluid and which will form vector fields defined throughout the whole sample. Given a container and external fields, **l** and **d** will form a configuration such as to minimise the energy of the system. For this to occur several conditions, which usually all compete with each other, have to be optimised: (1) **l** must be perpendicular to any wall, (2) **d** and **l** want to be as straight as possible because bending costs energy; (3) **d** wants to be perpendicular to an external magnetic field because of the susceptibility anisotropy and finally **l** and **d** want to be parallel due to the nuclear dipole interaction in a ^{3}He–^{3}He pair. It is this very energy that leads to the unusual magnetic properties of superfluid ^{3}He, which become particularly evident in nuclear magnetic resonance (NMR) experiments. The bulk A phase hence shows not only a transverse resonance frequency which is strongly shifted away from the Larmor frequency but also a longitudinal resonance! By means of NMR experiments one can then indirectly obtain information about the **l** and **d** configuration in a sample, because **d** will oscillate in the potential provided by the **l**-vector, giving rise to particular resonances. Of all textures the coupled **l** and **d** fields can build up, those which are non-uniform and show textural defects and singularities are of particular interest. Quite generally the topics of defects and singularities in condensed matter physics and their classification has been of great theoretical interest for the past 2 years (see, for example, Toulouze & Kleman, *J. Phys. Lett. (Paris)* **37**, L-149; 1976).

While some experimental work on non-uniform textures (without singularities) has been done already the first experimental studies which specifically deal with singular textures have only recently been reported (Saunders *et al. Phys. Rev. Lett.* **40**, 1278; 1978; Gould & Lee, *Phys. Rev. Lett.* **41**, 967; 1978).

Dieter Vollhardt is at the Department of Physics, University of Southern California, Los Angeles.

0028–0836/78/0276—0325$01.00

© Macmillan Journals Ltd 1978

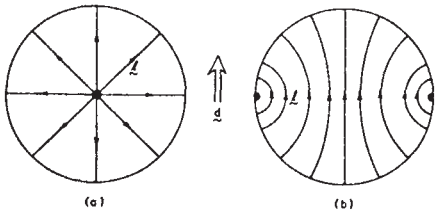


Fig. 1 Two possible planar textures of the **l** vector confined in a long cylinder. *a*, Modified de Gennes disgyration with the **l** vector strictly radial, giving a singularity in the centre, and the **d** vector uniform. *b*, The 'Pan-Am' texture with two singularities on the walls. **d** is again uniform (from Gould and Lee, *op. cit.*).

By a rare coincidence both groups carried out NMR measurements of A-phase textures in the same type of very narrow cylinders (2 μm diameter). In the experiments an external static magnetic field was applied parallel to the cylinder axes with rf fields parallel and transverse to this to measure the longitudinal and transverse resonances. In addition, Gould and Lee also reported measurements of a transverse resonance with the external magnetic field perpendicular to the cylinders. The confinement of the superfluid in these narrow cylinders has drastic consequences for possible **d** and **l** textures. As bending of **l** and **d** usually takes place over a distance of typically 10 μm, such a deformation will cost quite a lot of energy if it happens over a length of 2 μm, even more energy than to have a singularity somewhere in the plane of the cylinder. The only known non-singular texture that would fit into a cylinder, the so-called Mermin-Ho vortex, can be shown to require too much bending energy to be a favourable candidate, so that the only possible textures must bear singularities. With a magnetic field parallel to the cylinder one would thus *a priori* expect planar textures with **d** uniform across the sample (in the plane of the cylinder cross section) and the **l**-field looking somewhat as in Fig. 1.

Measuring the transverse NMR signal, both groups find two resonances where initially most of the weight lies in the higher frequency mode; however, this mode slowly decays into the lower mode over 40–60 min, which seems to imply a textural transition. The two groups measured somewhat different NMR frequencies and while the resonances of Gould and Lee proved to be temperature dependent (suggesting a temperature-dependent texture) the one of Saunders *et al.* did not depend on temperature. Their results for the longitudinal resonances are also different. Gould and Lee find a single, well defined, longitudinal resonance (temperature dependent) which does not change during the textural transition, while Saunders *et al.* do not observe any such resonance, which would be consistent with an axially symmetric texture. Whether the differences in the results can be explained by the different experimental setups (Gould and Lee for example, use Pomeranchuk cooling, Saunders *et al.* use nuclear demagnetisation) is unclear. The results, however, lead to the following conclusion: in both experiments the first texture seems to decay into a similar one, whose projection into the plane of the cylinder's cross section is about the same (so they give the same longitudinal resonance). The initial textures that would be fairly consistent with the NMR-resonances are the de Gennes disgyration in the case of Saunders *et al.* and a somewhat distorted 'Pan-Am' texture (Bruinsma & Maki, private communication) in the case of Gould and Lee. The deviations of the cylinder cross sections from a circular shape—they have a rounded hexagonal structure as can be seen on magnifications—probably need to be considered. The only alternative mechanism that could, at least qualitatively, account for a distortion of textures in the direction of the cylinder axes in both experiments and the temperature dependence in Gould and Lee's results seem to be superfluid currents in the cylinders.

To explain quantitatively what textures are responsible for the measured NMR resonances will not be an easy but most certainly a very rewarding task. It will provide us with new understanding of textures and singularities in superfluid ^{3}He, and beyond that in condensed matter systems. □

Germs from Mars?

from Arie J. Zuckerman

THE health and environmental hazards of bringing to Earth surface soil samples from Mars, assuming that life exists on that planet, have been discussed for over a decade. On the one hand, views have been expressed that the risk of the spread of virulent Martian organisms to terrestrial life forms is so great that such an undertaking planned for the 1980s should be abandoned (see for example Young & De Vincenzi *Science* **186**, 496; 1974). In a recent cogent discussion of the problem (*COSPAR: Life Sciences and Space Research* XVI (eds Holmquist & Strickland) Pergamon, 1978) Martin Favero points out that adequate containment facilities for microorganisms are now available, although absolute security in a scientific sense is of course impossible to guarantee. It is therefore necessary to examine the risks in terms of possibilities.

Favero develops his argument for the unlikelihood of Martian microorganisms' causing harm as follows. If life exists on Mars such organisms would be adapted to low temperature fluctuations and virtually no oxygen or water in the atmosphere. It could reasonably be assumed that such organisms will be so adversely affected by the Earth's environment that Herculean scientific efforts will be required simply to maintain them. Counter arguments could be that terrestrial microorganisms survive under wide ranges of conditions, which is in principle true, and furthermore Martian organisms might remain unrecognised by the host's immune and other defence mechanisms thereby resulting in devastating infections. However, in general terms, microorganisms that have no association with man are not pathogenic, for example hydrogen bacteria, methane bacteria and blue-green algae. Another argument is that extraterrestrial microorganisms might be resistant to inactivation, for instance to heat. It is pointed out that heat which destroys carbon bonds will destroy Martian organisms. If they survived, it would be because their biology was not based on carbon, in which case they should not be pathogenic for forms of life whose biology is based on a carbon system. Finally, there is the potential risk of exchange of genetic material between Martian organisms and terrestrial forms of life. However, there is no evidence of efficient transfer in nature of genetic material between heterogeneous broad groups such as bacteria, algae, fungi, protozoa, higer plants and animals, and the probability of introducing genetic material from Mars to Earth seems insignificant.

This is the stuff of science fiction. Yet similarly conjectural risks had to be considered in the development of atomic energy, production of antibiotics and the use of live virus vaccines.

Arie Zuckerman is Professor and Head of this Department of Medical Microbiology and the WHO Collaborating Centre for Reference and Research on Viral Hepatitis at the London School of Hygiene and Tropical Medicine.

0028–0836/78/0276—0326$01.00

© Macmillan Journals Ltd 1978

Allegory of Weather Mary Evans Picture Library

Sorry, for copyright reasons some images on this page may not be available online

climatology supplement

The World Climate Programme

B. J. Mason

IN recent years there has been an upsurge of interest in climate and its variations, the prediction of future trends and their consequences. Increasing world population and higher living standards have increased the pressure on natural resources of food, fibre, energy and water, the balance between the supply and demand for which could be seriously affected by even temporary changes in the climate of critical regions. The impact of any such changes will depend not only on their magnitude but also on the speed with which they occur. There is also increasing concern that man himself may induce inadvertent changes of climate by polluting the atmosphere with increasing quantities of carbon dioxide, heat, aerosols and chemicals. Although there has been a tendency to exaggerate the magnitude and consequences of the climatic fluctuations likely to occur in the next 100 years, future changes are likely to have greater economic and social impact than those of the past. The whole subject of climate, its variations and consequences therefore merits thorough and critical investigation.

In response, the World Meteorological Organization and the International Council of Scientific Unions, in collaboration with other international agencies, are planning a World Climate Programme spanning the two decades 1980–2000, preceded by a World Climate Conference in Geneva in February 1979 at which present knowledge will be reviewed and future plans discussed. The main objectives of the World Climate Programme may be summarised as follows: to improve knowledge of the natural variability of climate; to understand the underlying causes and mechanisms; to assess the sensitivity of climate to natural and man-made perturbations; and ultimately to predict future climatic changes; and to assist in the making of planning and investment decisions in climate-sensitive industries with a view to reducing the impact of adverse climatic changes. The Programme will have three main elements: a climate data and applications programme; a programme to investigate the impacts of climate on human activities; a programme for research on climatic change and variability.

Data and services

Although prediction of the magnitude and timing of climatic fluctuations is not yet possible, estimates of future probabilities can be made from analysis of past records of the main climatic parameters and these can serve as a useful input into the long-term planning of climate-sensitive industries such as agriculture, energy, water resources, transport and building and construction. A great deal of the available worldwide data are held in national meteorological services such as the British Meteorological Office where it is readily accessed by computer and used to answer some 30,000 enquiries each year from a wide spectrum of public services and industries. However, climatological observations need to be improved in many parts of the world where modern methods of data processing, analysis and presentation are required for the effective application of present knowledge to practical problems.

Impact studies

Very few studies have been made of the likely impact of marked and persistent changes in climate even on the major food-producing areas of the world. This would involve detailed analysis of relationships between weather and the growth and yield of the more important crops in different climatic environments. Such studies might pave the way for the forecasting of crop yields, on a regional or national basis, if and when reliable seasonal and longer term weather/climate forecasts become available.

Although it seems generally agreed that the greatest man-made potential threat to climate is the release of carbon dioxide by the burning of fossil fuels, no detailed and critical assessment of the likely effects has been made. It is not obvious that the overall result would be as detrimental as is usually assumed; it could even be beneficial. Although an average global warming of 5 °C with up to 10 °C in polar regions might well lead to substantial melting of the ice cap and widespread flooding, a more modest and more likely rise of 2–3 °C might well lead to increased food production by prolonging the growing seasons, increasing the rainfall and enhancing photosynthesis, and, furthermore, produce considerable savings in energy consumption. There will, no doubt, be some off-setting disadvantages but these should be assessed before assuming catastrophe.

Research

Study of past climates The reconstruction and study of past climates is necessary to determine the nature and extent of climatic fluctuations, to provide a guide to future changes and to validate computer simulations of past climates. Beyond the period of instrumental records the evidence rests strongly on modern techniques of radiocarbon dating and stable isotope analysis of fossils, ice cores, tree rings, and so on, which have greatly improved in recent years. These historical records, being dominated by irregular fluctuations, cannot be safely extrapolated to predict future changes. They can, however, be used to estimate the probability that fluctuations of a given magnitude, range and duration may occur within a given time interval and also to estimate the associated degrees of risk.

Climate monitoring Comprehensive data sets on the present climate are required to study mean annual and seasonal climates, for model development and validation, and to provide better climatological services. The First GARP Global Experiment (1978–79) will provide the first comprehensive set of weather observations covering one annual cycle. It will also be necessary to monitor at least the upper layers of the ocean, the ice sheets, changes in the land surface and its biological cover,

0028–0836/78/0276—0327$01.00

© Macmillan Journals Ltd 1978

the incoming solar radiation, and the carbon dioxide, ozone and dust content of the atmosphere. Most of these data, except subsurface ocean data, should be provided by satellites within the next few years but their processing and archiving will require considerable additional effort.

As it may not be possible to understand, much less predict, long-term climatic changes without treating the atmosphere, oceans, ice sheets, land surface and biosphere, as a single interactive system, the lack of oceanographic data may ultimately prove a serious obstacle. Although limited area and duration experiments, such as JASIN just concluded in the North Atlantic this summer, will help to elucidate interactions between the upper layers of the ocean and the atmosphere, long-term climatic changes may be induced by changes in the deep ocean circulation. These will be difficult to monitor on a fully global basis but carefully planned measurements in critical areas should provide important clues and will probably form part of the World Climate Programme to be implemented over several years as a concerted international effort.

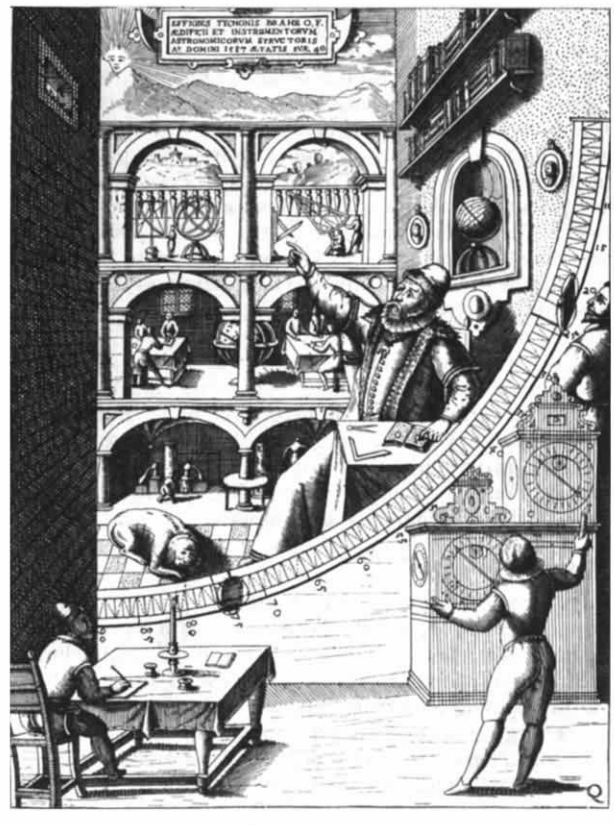
Computer modelling Although the processes involved in the maintenance of the Earth's present climate are broadly known, those responsible for climatic fluctuations are largely unknown. Predictions of the extent and duration of climatic changes are not possible at present and must await greater knowledge and understanding of the underlying causes, whether internal or external to the atmosphere. As the climatic system embraces the atmosphere, hydrosphere, cryosphere, lithosphere and biosphere and a great variety of processes and interactions over a wide spectrum of space and time scales, their synthesis requires the building of very complex computer models of the type described in this supplement by A. Gilchrist. However, the construction, testing and evaluation of such models is a major undertaking requiring scientific resources and great computing power that is presently available only in the USA and at the British Meteorological Office where remarkable progress has been made in simulating the main features of the present global climate including seasonal changes and important regional features such as the monsoons. The models may also be used with some confidence to investigate the response of climate to conceivable natural and man-made perturbations. For example, the British Meteorological Office (Mason, B. J. *Proc. R. Soc.*, **A**, in press) has carried out a wide range of model experiments to assess the likely effects of changes in solar radiation, soil moisture and vegetation cover, sea-surface temperatures, polar ice cover, and so on, and of conceivable man-made changes in the carbon dioxide, ozone, dust and heat content of the atmosphere, and to judge whether these are likely to be distinguishable from natural climatic fluctuations. Also, by testing the sensitivity of the statistics generated by the model to variations in a particular parameter or combination of parameters, one may hope to discover the underlying causes of climatic change. However, it is sometimes difficult to assess the significance and reliability of a particular result especially if the analysis of the experiment has not clearly revealed the response mechanism of the model and there have been insufficient tests to assess the significance of the signal relative to the noise generated by the model. The response (signal) to a prescribed perturbation will be statistically significant only if the difference in the mean values of the perturbed and standard (control) states can be discriminated from the model noise as estimated, for example, by its time-averaged response to random perturbations in the initial conditions. There is a general impression that the models exhibit less variance than the real atmosphere; if this is so, it may be easier to detect a perturbation signal above the noise in a model than in the real atmosphere. It therefore seems important to compare not only mean values but the models' simulated variance with observations. All this implies a long, arduous programme of model development, improvement and assessment with the object of narrowing the degree of uncertainty in the answers they provide.

Prediction The prediction of future climatic changes, as distinct from the simulation and sensitivity studies just mentioned, poses a formidable scientific problem partly because it is not possible to verify such long-term predictions against observation. Prediction of climatic variations induced by changes in external forcing factors such as the incoming solar radiation, changes at the boundaries of the climatic system, or by man's activities presupposes that these stimuli can be predicted accurately in advance. Aside from this the precision of any climatic prediction, expressed as a set of climatological statistics, will be limited by random or spontaneous internal fluctuations within the real climatic system and by the extent to which these are reproduced by the model. When integrated over long periods with fixed boundary and external conditions models are likely to approach an equilibrium state and produce climate statistics largely independent of the transient features of the initial state.

There is, however, the possibility that the real climatic system, being highly non-linear, may be able to exist in two or more quasi-stable modes and change rather abruptly from one such state to another. The transitions between glacial and interglacial epochs are suggestive of such changes. If such transitions are externally induced, they are probably predictable, at least in principle. If, however, they can arise spontaneously from internal reconfiguration of the climatic system they may not be predictable by models that are damped sufficiently to ensure dynamical and computational stability. In any case, soundly based long-range climatic predictions are not likely to be possible within the next decade.

I hope that this introduction may give some indication of the scope, complexity and difficulty of the climatic problem, illustrate the need for substantial and sustained programmes at both national and international level, and engender a more sceptical attitude towards the ill-founded and often alarmist statements that bedevil this subject. □

B. J. Mason is Director-General of the Meteorological Office, Bracknell, UK.



Tycho Brahe in seiner Sternwarte „Uranienburg" auf der Insel Hven
Nach einem Kupferstich zu Tycho Brahes „Astronomiae instauratae Mechanica" vom Jahre 1602

One of the 'earliest climatologists', Tycho Brahe (1546–1601), in his observatory on the island of Hveen (now Ven, Denmark) (taken from the 1602 edition of his *Astronomiae Instauratae Mechanica*)

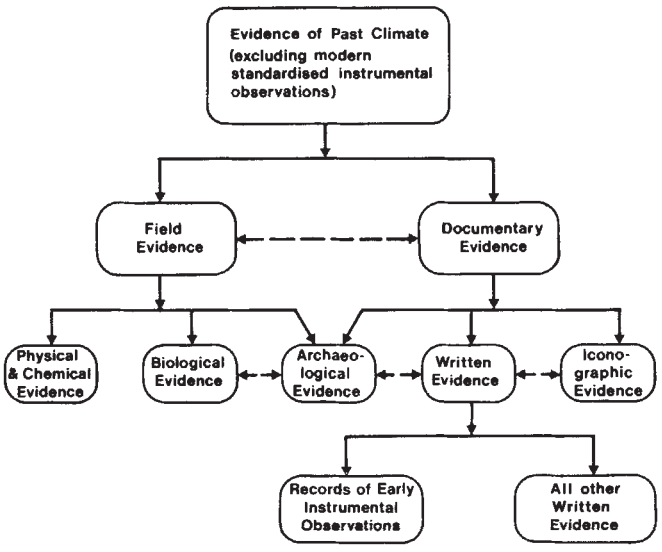
Historical climatology

M. J. Ingram, D. J. Underhill & T. M. L. Wigley
Climatic Research Unit, School of Environmental Sciences, University of East Anglia, Norwich, UK

Descriptive documentary evidence is an important source of detailed information on past climates, particularly for the period between the eleventh century and the beginning of the era of instrumental meteorology. Historical climatology is concerned with the study and climatic interpretation of this evidence. However, there are many pitfalls in using historical records, associated with the fact that such records were rarely written primarily as descriptions of climate and with the fact that many readily accessible records are of doubtful reliability. Here we take a critical look at existing work in the field, with the aim of isolating many problems which need to be recognised and overcome before historical climatology can be considered to rest on a secure methodological and factual footing.

HISTORICAL CLIMATOLOGY emerged as a discipline in the late nineteenth and early twentieth century, the first major work of synthesis being by Brooks[1]. The subject's existence was at first fitful, but in the 1940s, 50s and 60s, the enthusiasm of men like Flohn, Manley, Schove, and above all, Lamb, successfully demonstrated the interest and importance of this field of study. In recent years, interest in historical climatology has grown considerably, and significant contributions have been made not only by natural scientists but also by historians (especially Ladurie[2]). This interest has arisen through an increased awareness of man's susceptibility to the vagaries of climate (the Russian wheat failure of 1972, and the Sahelian drought come readily to mind), through the realisation that climatic conditions very different from today have occurred in the recent past, and through the maxim of the acturial approach to climatic forecasting, "the past is the key to the future". In spite of this progress, no comprehensive methodology of historical climatology has been established, with the result that there exists no ready means for assessing the significance of individual contributions and, thus, of what has been achieved in the field as a whole. In view of this fact, we believe that the most useful method of reviewing the subject is to examine the problems which need to be universally recognised and overcome before historical climatology may be said to rest on a firm methodological and factual footing. In so doing, we shall as far as space permits criticise and evaluate past and current work in the field.

Fig. 1 Flow chart showing data useful for reconstructing past climates and delineating the contribution of historical climatology. The separate categories often overlap, and some of the more important interactions are shown.



Scope of historical climatology

The task of reconstructing the climates of the past involves the use of many different types of evidence (Fig. 1). The successful exploitation of this material demands a varied range of skills and techniques which effectively define specialised subdisciplines of climatology. Historical climatology is best thought of as one such subdiscipline, which focuses on the study of written materials (excluding records of modern standardised instrumental observations) which bear on past climate. These materials include, not only meteorological information, but also data on such phenomena as glacier movements, phenological events and other more or less indirect indicators of climatic change.

The distinctive contribution of historical climatology ends with the advent of instrumental observations, embracing pressure and temperature measurements, which effectively began in the mid-seventeenth century. The earliest series relate to Europe. In most other parts of the world, instrumental data extend back no further than the late eighteenth century, and in many locations span less than 100 years. The earliest instrumental series, moreover, pose many problems of interpretation, and the precise dates from which they may be regarded as acceptably accurate and standardised are debatable. The elucidation of early instrumental data may be regarded as within the scope of historical climatology, but the techniques involved[3-5] are specialised ones and will not be discussed here.

Apart from early instrumental records, the written sources of climatic data are legion. They include: (1) ancient inscriptions; (2) annals, chronicles, and so on); (3) records of public administration and government, whether imperial, national, provincial or local; (4) private estate records; (5) maritime and commercial records; (6) personal papers (diaries, correspondence, and so on); (7) scientific and proto-scientific writings (including journals of non-instrumental weather observations).

How far back into the past can the historical climatologist penetrate with the aid of these records? The answer varies with location. It is impossible to give a comprehensive guide to the earliest dates from which materials exist in different areas, because much essential exploratory work to discover possible sources has not yet been carried out. For example, it seems likely that the written remains of the brilliant and complex cultures of the Middle East and India would provide rich materials for the historical climatologist, but at present they remain almost totally unexploited. However, Table 1 provides enough information to indicate the main patterns of distribution. It can be seen that in certain areas the vision of the historical climatologist is limited to only a few decades preceding the establishment of standardised instrument observations. But for considerable areas of the world there are materials extending over centuries, and in favoured locations over thousands of years.

Even in areas for which early materials survive, the distribution of sources is by no means uniform over time. For example, the European evidence for the period before about AD 1000 is very scanty. It is appreciably richer, but still very

© Macmillan Journals Ltd 1978

Table 1 Earliest dates from which written evidence of climatic phenomena survives in selected areas

Area	Earliest written evidence (approximate dates)
Egypt	3000 BC
China	2500 BC
Southern Europe	500 BC
Northern Europe	0
Japan	AD 500
Iceland	AD 1000
North America	AD 1500
South America	AD 1550
Australia	AD 1800

imperfect, for the period AD 1000–1600. Thereafter many more materials are available, but it is only in the late seventeenth and eighteenth centuries that sources become really abundant. Equally, there are great differences in quality between different types of materials, with nearly all the best sources relating to the most recent periods.

As the sources are of such variable quality and quantity, a rigorous methodology is required to reject unsuitable and inadequate material and to extract the maximum amount of information from what remains. In the following, a methodology is discussed with particular reference to data verification and analysis of information.

Data verification

Written statements about the past are not equally reliable, and indeed much material which purports to record historical events is seriously misleading. Historians have developed a sophisticated methodology for evaluating sources and rejecting unreliable information. Plainly, historical climatologists need to be similarly aware, both of the need for source criticism, and of the available critical techniques. Scientists attempting to reconstruct past climates have, in fact, often been ignorant on one or both of these counts (see, for example, criticisms in the works of Gottschalk[6–8], Alexandre[9], and Bell and Ogilvie[10]).

A critical discussion of the sources of data available to historical climatologists will reveal the significance of this fact. There exist many eighteenth-, nineteenth-, and twentieth-century compilations and syntheses of historical climatic information (the most important are listed by Bell and Ogilvie[10]). At first sight, these works seem to provide a conveniently ready-made data bank, and it is therefore not surprising that they have formed the basis of numerous attempts at climatic reconstruction. In reality many of these compilations and syntheses are highly defective as data sources, and work founded on them is liable to be inaccurate. Their deficiencies may be summarised as follows. First, their transmission of the original data is often inaccurate and incomplete. Compilations and syntheses often consist of abridged, translated, paraphrased, or condensed material. Second, many of the works under discussion include errors introduced by the author or compiler, the most common mistake being misdating. Third and most fundamentally, most compilations fail to distinguish adequately (if at all) between reliable and unreliable sources. It is true that some very recent works[6–8,11,12] are models of critical awareness. As a whole, however, the available compilations and syntheses comprise a mishmash of valuable and worthless data, and to use them uncritically for the purposes of climatic reconstruction is unacceptable. An illustration of the amount of valueless material contained in two widely used compilations is given in Table 2. In constructing this Table we have subjected the compilations of Hennig[13] and Weikinn[14] to a critical historical analysis (see below) and counted the number of items which survive this criticism. Such items are described as 'verified'.

When scientists attempt to penetrate beyond compilations and modern syntheses to the materials on which they are based, or if they seek new data, they are faced with the problem of distinguishing between reliable and unreliable sources. Various criteria may be applied to test reliability. The tests are numerous and complex, and here only the basic principles are indicated. Most fundamentally, it is necessary to examine the nature of the source to assess whether, or how far, the writer was actually interested in recording real events, for not everything which looks like history as we understand it was intended as such. Of further tests, the most important are those based on the principles of contemporaneity, propinquity, and faithful transmission. Recorded statements cannot be regarded as reliable unless it can be shown either (1) that the writer lived close in time and space to the events he purports to describe, and recorded his observations either immediately or within a very short time after these events had taken place, or (2) that he had access to first-hand oral or written reports and accurately transmitted the information derived from them.

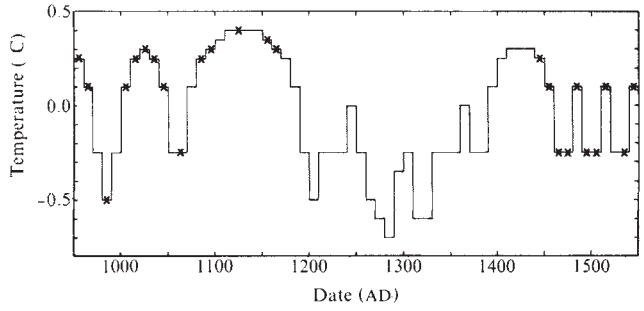


Fig. 2 Bergthórsson's original decadal estimates of temperature in Iceland showing the sparseness of data which remain after a critical examination of the historical evidence. The decades for which there is no historically acceptable information are marked with crosses. (Information supplied by Astrid Ogilvie.)

Many materials, especially narrative sources such as chronicles which were often subject to constant addition and revision, emerge badly from such tests. It is frequently found that sources are in whole or in part either (1) slavishly derivative and therefore of no independent value; (2) inaccurately and corruptly derivative; (3) spurious, fabricated for some particular

Table 2 Historical reliability of two European weather compilations for the period AD 0–1000

Century (AD)	Hennig[13] Total no. of seasons described	Hennig[13] % Of seasons verified	Hennig[13] Additional information by season	Weikinn[14] Total no. of seasons described	Weikinn[14] % Of seasons verified	Weikinn[14] Additional information by season	Total no. of reliable seasonal descriptions available
0–99	20	50	2	8	88	5	12
100–199	15	7	2	2	50	2	3
200–299	9	0	1	1	100	—	1
300–399	16	25	3	7	86	1	7
400–499	22	14	—	2	100	1	3
500–599	44	25	5	14	71	6	16
600–699	23	13	1	3	0	4	4
700–799	31	13	6	9	56	5	10
800–899	47	57	23	35	91	18	50
900–999	45	29	6	8	75	13	19
Total	272	28	49	89	79	55	125

The additional information' refers to reliable material which we have extracted from other primary sources.

purpose such as to assert the rights and claims of the author's monastery; or (4) of literary, philosophical, or religious significance but of no value as records of events.

Much of the data which historical climatologists have hitherto used have been drawn, either directly or more often by way of compilations, from such unreliable sources, and are therefore subject to error. The most important errors are: (1) inaccurate or uncertain dating of particular events; (2) spurious multiplication of events through failure to recognise that accounts recorded under different dates in various sources in fact relate to a single event; (3) acceptance of accounts which are distortions or amplifications of original observations (for example, purely local events may appear through the distortions of later writers as major catastrophes covering a large area); (4) inclusion of events for which there is no reliable evidence whatsoever (such 'non-events', moreover, tend to be artificially multiplied through being assigned to different dates in subsequent derivative accounts).

These errors are not randomly distributed and are therefore not self-correcting. On the contrary, the problem of unreliable data is so great that decadal or even 50-yr series are liable to be grossly inaccurate if unverified material is used. These points may be illustrated by reference to the work of Bergthórsson[15-17] and Lamb[18,19], although it must be stressed that the studies in question were onlh presented as tentative preliminary analyses. Figure 2 shows the shortcomings of the early part of Bergthórsson's decadal temperature series for Iceland, itself based on the work of Thoroddsen[19]. The decades marked with a cross are those for which, as critical analysis reveals, no reliable climatic information is available. Only eight of the first 22 decades of the record have any reliable climatic information, and many of these decades have only one or two items. For the period AD 1431–1560 only one item of reliable climatic information is available. Clearly it is quite impossible to construct decadal temperature estimates for much of the period, and even 50-yr averages cannot be estimated before AD 1100 or between AD 1450 and 1550. Bergthórsson himself may have had misgivings about the decadal series shown here as it has not been published in its original form: nevertheless, it has subsequently been reproduced by other authors[15,16].

Lamb[9] has published decadal indices for England, Central Europe and the Russian Plain for the period AD 1100 onwards (discussed further in the next section). As with the Icelandic reconstruction, these indices are partly based on compilations which are now known to be unreliable, and, on this ground alone, they must be regarded as suspect: they are currently being revised in the Climatic Research Unit. Although the results of this revision are not yet complete, it is instructive to compare Lamb's indices with the work of the historians Titow[21] and Alexandre[11,12].

Fig. 3 Comparison of winter temperatures (December, January, February; °C) from three European sites for the period 1684–1738. Short-term fluctuations have been smoothed out by using a 9-yr Binomial filter. The Zürich record[30] is based partly on the observed correlation between Zürich and central England temperatures, but also on records of local snowfall and duration of snow cover. (We have filled some minor gaps in the Zürich record using the Zürich–England correlation.) Part of the central England record[31] (1707–22) is based on instrumental data from Delft.

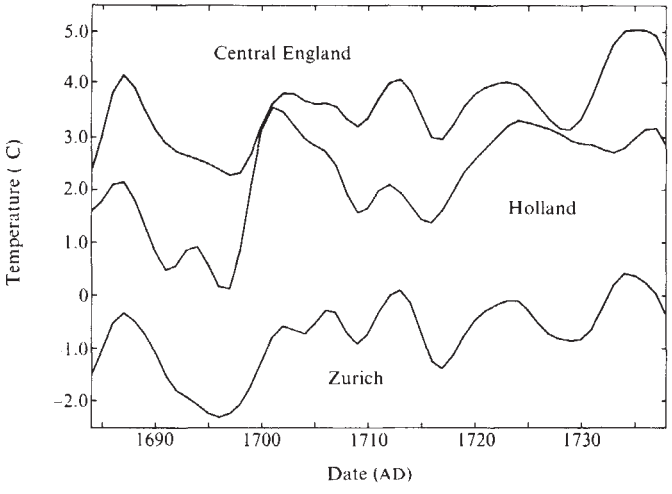


Fig. 4 A frost fair on the River Thames early in 1863.

Consider first the summer wetness indices for England for the period AD 1200–1350. Lamb (ref. 18, page 32) has compared his results for these 15 decades with primary data extracted by Titow[21] from the manorial account rolls of Winchester Cathedral. The two records are well correlated ($r = 0.77$). On the other hand, Lamb's English winter severity indices for the period AD 1100–1400 seem to be less well-founded. Here, there is no independent record for England, but there is a similar record available for Belgium constructed by Alexandre. (Alexandre[11,12] has analysed the compilation of Vanderlinden[22] in detail and found it to be seriously unreliable. In order to fill the data gaps which resulted from his source criticism, Alexandre examined a large number of primary sources which Vanderlinden had not used. From this new body of data he then calculated decadal indices using the method developed by Lamb (see below).) Comparison of Lamb's and Alexandre's indices shows only a low correlation ($r = 0.2$, compared with a correlation coefficient of 0.9 for decadal winter temperatures between AD 1750 and 1940). Although this low correlation may be partly the result of real changes in meteorological factors, it must be largely attributed to the variable, but generally poor quality of the compilations used by Lamb.

These strikingly contrasting results indicate the need for a re-examination of the data which form the basis for the currently accepted historical climate record. It is essential to build up a data bank of historical information of unquestionable reliability.

Analysis and interpretation

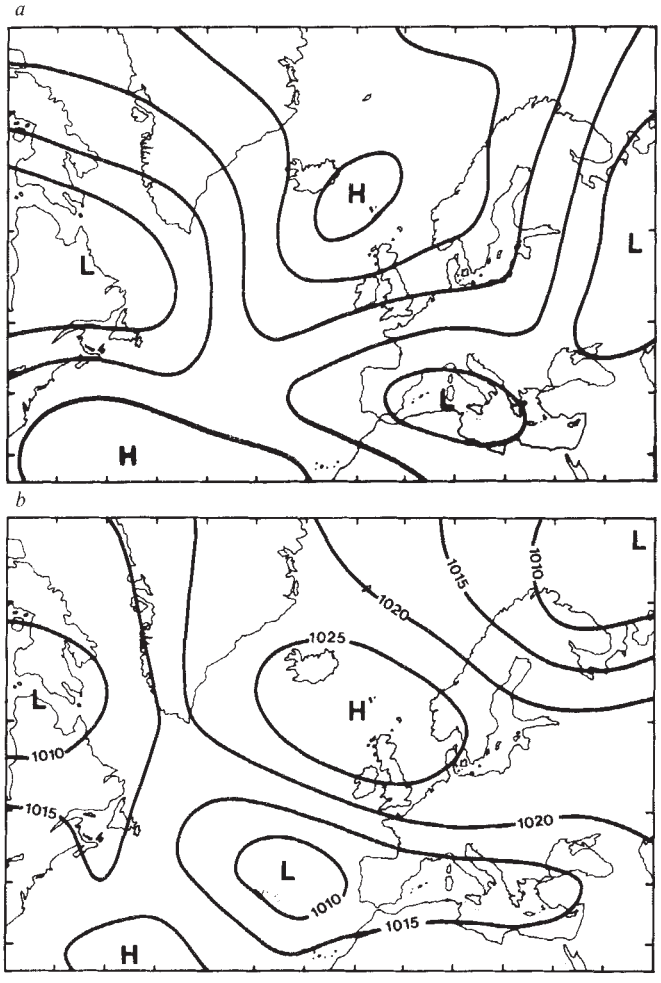
Given reliable data, the next task of the historical climatologist is to use different types of documentary evidence to produce time series which are, as far as possible, comparable with modern meteorological records. Four steps are involved: historical analysis of the material; quantification; transformation into meteorological variables; and the addition of long-term trends where necessary.

Modern instrumental observations provide the nearest approach to an objective and comprehensive record of climatic phenomena. Historical climatic data, on the other hand, embody distortions and omissions reflecting the various personal and cultural biases of past observers. The ways in which phenomena were perceived and categorised varied from period to period, place to place, and observer to observer. So also did the principles by which phenomena were selected for the record. Even the manner of recording—terse or verbose, sparsely or profusely detailed, matter-of-fact or poetically embellished—affects the information content of the record.

In the final analysis, these factors limit the quality and quantity of the information which can be derived from the documents. But they also pose problems of interpretation, in effect obscuring what information the materials do contain. This obscurity must, as far as possible, be removed by subjecting the sources to rigorous textual analysis to establish the meanings of words and categories and to determine the nature of biases. This may be achieved by reference to specialised historical knowledge of the intellectual and social milieux which gave rise to the texts, and by internal analysis of individual sources and comparison between sources relating to the same set of phenomena. Moreover, as Moodie and Catchpole[23,24] have demonstrated, the methods of analysis should be codified and systematised to ensure that individual analysts, faced with the same body of materials, will derive essentially the same information.

Although a universal set of rules cannot be defined, the analyst must ask of the verified data these basic questions: (1) Why was the information recorded? How did the purposes of the writer bias his perception and selection of phenomena? (2) What is the precise significance of such categories as 'frost', 'drought', 'flood', 'sea-ice', and so on? (3) What is the significance of terms of degree? For example, what weight should be attached to expressions like "The greatest . . . that hath been known in the memory of any man living".? (4) When did the event occur, and for how long? Changes in the calendar have occurred at different times in different places, and the precise timing of an event is not always important to the past recorder. Furthermore, the time-span covered by terms like 'summer' and 'winter' varies from record to record. (5) Where did the event occur? Did the recorder distinguish between local and 'foreign' events? The weather diary of Merle[25], kept variously in Lincolnshire or Oxford from 1337–44, is a mild example of the sort of confusion which can arise.

Fig. 5 Preliminary reconstruction of the winter 1683–84 circulation pattern (*a*), and the pressure map for January 1963 (*b*). (1683–84 reconstruction supplied by H. H. Lamb.) Both seasons exhibited strong blocking patterns with persistent E to NE winds over England. The normal winter pattern shows a belt of low pressure at the latitude of Iceland, and south-westerly winds over England, and the patterns here are highly anomalous.



The next stage in analysis is to quantify the information in terms of standard meteorological variables. In general, two steps are necessary: quantification in terms of a numerical index followed by calibration of the index in terms of a variable such as temperature or precipitation. Certain types of information are already numerical in form, and if they are homogeneous and systematic they can be calibrated directly by correlation between the original record (or a modern-day equivalent) and a meteorological parameter provided there is a sufficient period of overlap between the records. One of the earliest examples of this type of analysis is the winter temperature reconstruction derived by Flohn[26] from the meteorological journals of Tycho Brahe[27] (Hven, Denmark, 1582–97) and Haller[28] (Zürich, 1546–76). Flohn found from modern instrumental observations that the ratio of days of snowfall to days of rainfall in winter was well correlated with winter temperature. Using the snow and rain records of Brahe and Haller, Flohn was able to show that winters in Zürich were about 1.3 °C colder in 1561–76 than in 1546–61, whereas the winters on Hven from 1582–97 were about 1.5 °C below those around the turn of the twentieth century. The implied severity of the latter part of the sixteenth century accords with the wind direction observations given by Brahe and analysed as early as 1876 by La Cour[27].

Similarly, De Vries[29] has calibrated data on freezing of water in Dutch canals in terms of winter temperatures to extend the record of Dutch winter temperatures back to 1634; and Pfister[30] has used data on snow falls and duration of snow cover from various historical sources to estimate winter temperatures near Zürich back to 1683. Figure 3 compares De Vries' and Pfisters' results with Manley's[31] central England temperature record (based on instrumental data) for the period 1684–1738. All three records show the striking cold of the 1690s, a decade without subsequent parallel.

Unfortunately, most historical information is not so readily quantifiable in terms of meteorological variables, as the data are generally discontinuous, non-homogeneous, and show a marked bias towards the recording of extreme events. In such cases the documentary evidence must first be transformed into a numerical index. One of the first to do this was Brooks[1] who derived 50-yr indices of wetness and winter severity for Europe using data from a number of compilations. Lamb[19] refined this work by calculating decadal indices for Europe near 50°N at three longitudes, 0°E (England), 25°E (Central Europe) and 50°E (the Russian Plain), back to AD 1100. Lamb's indices may be defined as follows. For 'winter severity', the index is $C-M$, where C is the number of unmistakably cold winter months in a decade and M the number of unmistakably mild months; and for 'summer wetness', $W-D$, where W is the number of summer months with evidence of frequent rains and D is the number of months with evidence of drought per decade.

One of the underlying assumptions of indexing, and the basis for expecting indices to reflect the meteorological variables, temperature and precipitation, is that there are few, if any, severe seasons for which we have no written record. Previous historical climatologists have tended to believe that there are few such data gaps, mainly because they had no way of knowing the poor quality of their sources. We now know that there are gaps in the record, especially before AD 1500, and that early impressions were false because of the amount of duplicated, fabricated and erroneous information in the widely used compilations.

This raises a question: how many gaps are permissible before decadal index values become unrepresentative? Different types of index show differing sensitivity to missing data. 'Difference' indices, as used by Lamb, are quite sensitive, whereas 'ratio' indices show less sensitivity. (An example of a ratio index for winter severity is $(C-M)/(C+M)$ where C and M are the

number of cold and mild months, respectively.) The sensitivity of indices can be tested by a Monte Carlo type of simulation where one starts with an artificially constructed (but realistic) complete record and progressively introduces more and more random gaps until there is a significant change in the average index value for an ensemble of decades. Such simulations show that ratio indices can tolerate small gaps in the record. (Equally important, they indicate that ratio indices are robust to uncertainties in the perception and/or actual frequency of extremes.)

The second step in quantification is to convert from index to meteorological parameter. Here the method is to correlate the index values with instrumental temperature or precipitation data over an overlap period. If a significant correlation cannot be established then the index itself is of doubtful value, but if there is a correlation then it can be used to interpret index values before the overlap period directly in terms of meteorological variables. Lamb[32] found that the correlation coefficient between summer wetness and late summer (July plus August) precipitation was 0.91, and the correlation between winter severity and mean winter temperature (December, January, February) was −0.87. However, neither index shows any significant correlation with summer temperatures. We note again hat extrapolation back in time, even using the above high correlations, is suspect because, whereas the index values are reliable for the correlation period, their reliability is increasingly uncertain earlier in the record.

Courtesy Trustees, National Gallery, London

Sorry, for copyright reasons some images on this page may not be available online

Detail from *A Winter Scene with Skaters near a Castle* (about 1609) by Hendrick Avercamp (Netherlands)

The third step in producing meteorological time series from historical data is to introduce long-term fluctuations. It has already been noted that historical data (excluding the most easily quantifiable items discussed initially) tend to be biased towards the reporting of extreme events. But 'extreme' here is not absolute, and is only relative to the observer's own idea of what is normal, which in turn is largely determined by his own experience. Long-term trends in the norm go unobserved, particularly because they are small relative to year to year fluctuations. As a crude meteorological instrument, man tends to act as a high pass filter and his subjective observations show only short-term fluctuations about an ever-changing norm. The response of this filter is uncertain: for example, a trend may show up as a series of extreme years giving the appearance of a step-function change in climate.

Courtesy Trustees, National Gallery, London

Sorry, for copyright reasons some images on this page may not be available online

Detail from *The Castle of Muiden in Winter* (1658) by Van Beerstraaten (Netherlands)

This problem has long been realised by climatologists (see, for example, Brooks (ref. 1 pp. 329–330) and the solution suggested by Brooks still applies. This is to use either proxy data, or records which are not subject to filtering, to derive long-term trends independently, and to use these as the background for shorter-term fluctuations. Glacier fluctuations, pollen records, lake levels, tree line changes, vegetation limits and others are suitable. Some of these items may themselves derive from historical documents and so require care in compilation and analysis, whereas others are imperfectly understood as climatic proxies and can only be used after detailed individual studies of their responses to climate.

A more ambitious use of historical data is to construct a time series of maps of seasonal circulation patterns based on qualitative data from different locations. Such maps provide a synthesis of the diverse types of data which are available for Europe and the North Atlantic. Starting with a well distributed set of climatic descriptions pertaining to a particular season (or month), putative isobars can be drawn so that the implied wind pattern and the positions of centres of cyclonic and anticyclonic activity are consistent with the descriptive evidence. The principle involved is the same as that used by twentieth-century meteorologists to extend the surface pressure patterns on synoptic maps outside the region for which pressure data are available. Wind direction data are particularly useful for this type of analysis. The results depend on the experience of the analyst whose performance can be tested by using descriptive material from times when instrumental pressure maps are available. An example of this work is shown in Fig. 5, a reconstruction for the severe winter of 1683–84 (see also Fig. 4), arguably the most severe winter in England since 1659, and one of the most severe for Europe as a whole (possibly surpassed by 1708–09; ref. 33). The pressure map for January 1963 is shown for comparison: the similarity is

Mary Evans Picture Library

Sorry, for copyright reasons some images on this page may not be available online

Dutch winter scene (seventeenth century engraving)

striking and reinforces earlier comments by Manley[34] on these two winters.

Using the reliable historical information which is now available, it should be possible to reconstruct circulation maps for many of the extreme seasons in Europe back to at least AD 1500, covering the worst period of the 'Little Ice Age'. The Little Ice Age is a period which is unique in the record of the past 3,000 years at least[35]. Most importantly, it has no equal in the period of instrumental meteorology, and may provide information on 'no analogue' situations which might recur in the future. But is this period just one where extreme events similar to those of today occurred more frequently; or were the extremes of the seventeenth century in some way fundamentally different from those of today? This type of analysis may help us to answer this question.

Conclusions

We have tried here to show a few of the shortcomings of existing research in historical climatology, tempered by some of its successes, and its potential for the future. The main defect in previous work is that much of it can now be seen to have been based on the sands of unreliable data. We have demonstrated the significance of this in two ways: by analysing some of the commonly used sources to find out the extent of unreliability (see Table 2), and by comparing some accepted reconstructions with more historically acceptable analyses. We believe that, at the very least, the details of shorter time scale fluctuations are subject to considerable doubt, and that revisions are essential. Nevertheless, although many of the studies published to date are open to criticism, we would emphasise that their authors were the pioneers in a developing science, whose efforts have aroused scientific and public interest. The most appropriate way of appreciating their work is to develop more rigorous methods through the application of established historical techniques and the refinement of current methods of quantification. We believe that this will result in a significant improvement in man's knowledge of past climates.

We thank Astrid Ogilvie for access to her verified data for Iceland, and H. H. Lamb for the 1683–84 reconstruction used in Fig. 5. This work was supported by grants from the Rockefeller Foundation (M. J. I. and T. M. L. W.) and the US Department of Commerce (D. J. Underhill), the latter through the US National Oceanic and Atmospheric Administration. We thank J. Craddock, H. H. Lamb and G. Manley for useful discussions.

1. Brooks, C. E. P. *Climate Through the Ages* (Benn, London, 1926).
2. Le Roy Ladurie, E. *Times of Feast, Times of Famine: A History of Climate since the Year 1000* (Allen and Unwin, London, 1972); originally published as *Histoire du climat depuis l'an mil* (Flammarion, Paris, 1967).
3. Craddock, J. *Q. Jl R. met. Soc.* **102,** 823 (1976).
4. Lamb, H. H. & Johnson, A. I. *Geophys. Mem. 110* (HMSO, London, 1966).
5. Manley, G. *Q. Jl R. met. Soc.* **79,** 242 (1953).
6. Gottschalk, M. K. E. in *Stormvloeden en rivieroverstromingen in Nederland. Deel I: De periode voor 1400* (Van Gorcum, Assen, 1971).
7. Gottschalk, M. K. E. *Stromvloeden en rivieroverstormingen in Nederland. Deel II: De periode 1400–1600* (Van Gorcum, Assen, 1975).
8. Gottschalk, M. K. E. *Stormvloeden en rivieroverstromingen in Nederland. Deel III: De periode 1600–1700* (Van Gorcum, Assen, 1977).
9. Alexandre, P. *Le Moyen Âge* **80,** 101 (1974).
10. Bell, W. T. & Ogilvie, A. E. *Climatic Change* **3** (in the press).
11. Alexandre, P. *Centre Belge d'Histoire Rurale*, Publ. no. 50, (1976).
12. Alexandre, P. *Annales: Économies, Sociétés, Civilizations* **32,** 183 (1977).
13. Hennig, R. *Abh. Preuss. Inst.* Band II, No. 4 (1904).
14. Weikinn, C. *Quellentexte zur Witterungsgeschichte Europas von der Zeitwende bis zum Jahre 1850: Hydrographie*, Teil 1 (Berlin Institut für Physikalische Hydrographie, Berlin, 1958.)
15. Bryson, R. A. *Science* **184,** 753 (1974).
16. Karlén, W. & Denton, G. H. *Boreas* **5,** 25 (1976).
17. Bergthórsson, P. *Jökull* **19,** 94 (1969).
18. Lamb, H. H. *Climate: Present, Past and Future* **2** (Methuen, London, 1977).
19. Lamb, H. H. & Johnson, A. I. *Geogr. Annaler* **43,** 363 (1961).
20. Thoroddsen, Th. *Climate of Iceland in 1000 Years* (Hinu Íslenska Fraeðafjelagi, Copenhagen, 1916/17).
21. Titow, J. Z. *Annales: Economies, Sociétés, Civilizations* **25,** 311 (1970).
22. Vanderlinden, E. *Mém. Acad. r. Belg. Cl. Sci.* Ser. 2, **5,** (1924).
23. Moodie, D. W. & Catchpole, A. J. W. *Manitoba geogr. Stud.* **5** (1975).
24. Moodie, D. W. & Catchpole, A. J. W. *Science* **193,** 51 (1976).
25. Symons, G. J. *Considerationes Temperici pro 7 Annis: The Earliest Known Journal of the Weather, kept by the Rev. W. Merle, Rector of Driby, Lincolnshire, 1337–44* (reproduced and translated, 1891).
26. Flohn, H. *Vjschr. naturf. Ges., Zürich* **94,** 28 (1949).
27. LaCour, P. *Tyge Brahes Meteorologiske Dagbog, holdt paa Uranienborg for Aarene 1582–1597* (Kgl. Danske Videnskabernes Selskab: Collectanea Danica, Copenhagen, 1876).
28. Wolf, R. *Met. Beob. dt. Feuerschiffen N.- u. Ostsee* Suppl. 180, 1885).
29. De Vries, J. *Annales: Economies, Sociétés, Civilizations* **32,** 198 (1977).
30. Pfister, C. *Vjschr. naturf. Ges. Zürich* **122,** 447 (1977).
31. Manley, G. *Q. Jl R. met. Soc.* **100,** 389 (1974).
32. Lamb, H. H. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **1,** 13 (1965).
33. Lenke, W. *Ber. dt. Wetterd. Bad Kissingen* No. 92 (1964).
34. Manley, G. *Weather* **30,** 381 (1975).
35. Wigley, T. M. L., Underhill, D. J., Williams, L. D., Briffa, K. R. & Harris, P. W. V. *Quatern. Res.* (in the press).

Tree-ring evidence of past climatic variability

Valmore C. LaMarche, Jr

Laboratory of Tree-Ring Research, University of Arizona, Tucson, Arizona 85721

The increasingly visible impact of climatic variability on human affairs lends a sense of urgency to the task of better understanding the workings of the Earth's climatic system. Actual instrumental observations of climate are relatively short, and we must therefore turn to other sources for information about past climates to help develop and test the models that may enable us to predict climatic anomalies such as prolonged droughts or a series of severe winters. Tree-rings are one of the best sources of climatic proxy information. They can provide long, accurately dated, year-by-year records at many points around the globe, and bridge the gap between recent instrumental or historical data and the longer but more generalised geological records. Variations in the width of annual rings reflect the influence of climatic factors that limit the biological processes governing ring formation within a tree. Study of reconstructions of long records of a variety of climatic and related variables, such as temperature, precipitation, stream runoff and barometric pressure over periods of several hundred to several thousand years strongly suggests that the climate of the past century or so is not representative of the conditions that have frequently prevailed over long periods. Proxy records are thus a great help in our efforts to anticipate or predict future climate, which may be significantly different from the recent climatic past.

GREATER understanding of the global climatic system has become an increasingly important goal of scientific endeavour in a world that is faced with an expanding population and increasing pressure on energy, water and food resources. Growing awareness of the problem of climatic variability and its impact on human affairs is demonstrated by a series of recent major reports by the US National Academy of Sciences[1–3], the Australian Academy of Science[4], the Science Council of

© Macmillan Journals Ltd 1978

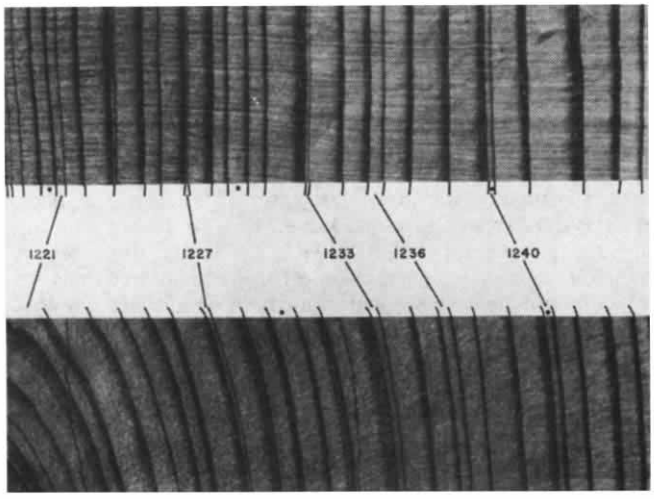


Fig. 1 Climatic sensitivity and crossdating. Douglas fir (*Pseudotsuga menziesii*) beam from a late thirteenth century Indian ruin in northern Arizona shows distinctive sequences of wide and narrow rings reflecting rainfall variations. The narrow rings represent years that were drier than normal.

Canada[5], and by other national and international agencies.

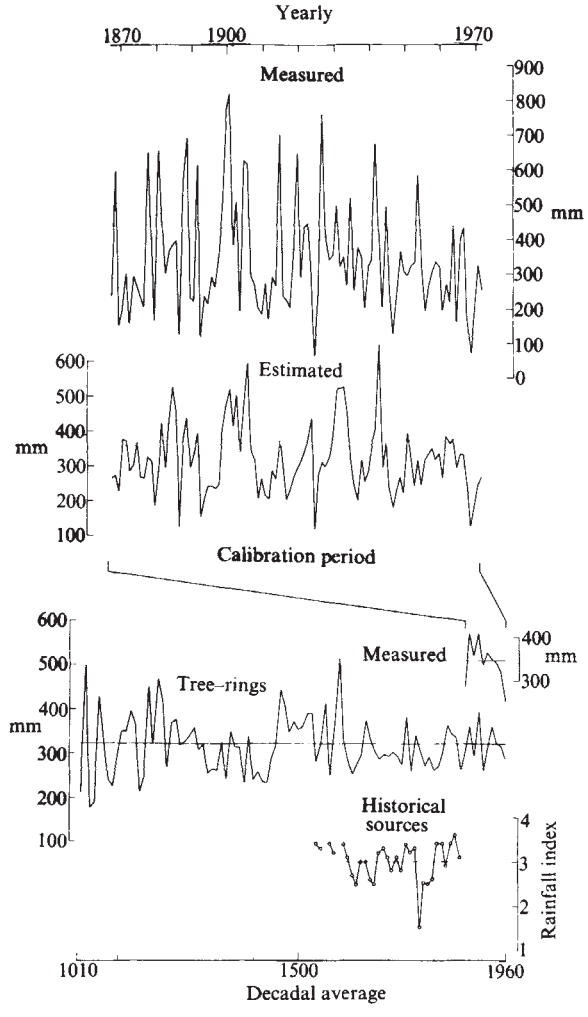
The basic problem is that we do not know enough about the workings of the climatic system, which includes the oceans, land areas, and ice sheets as well as the atmosphere. Because of its complexity, predictions about the future behaviour of the system must ultimately come from physically based numerical models which embody our understanding of the processes involved. However, the existing observational data base may be inadequate for the development and verification of models capable of accurate climatic forecasts. One of the most serious constraints that could emerge is that a model 'tuned' to generate the present climate with adequate precision may not be able to reproduce climatic states that have existed in the geologically recent past, and that could recur in the immediate future. An important step in the verification of a model (with important implications for its use as a forecasting tool) would be to test whether it can generate not only plausible mean conditions, but also time series with variance spectra similar to those of data from the real world. Observational, historical and climatic proxy data for time periods independent of those used in development are thus needed for complete model verification.

Palaeoclimatic records

The records of actual instrumental observations of climate are inadequate for study of climatic variations on time scales of decades to millennia[1]. Continuous records of surface elements such as atmospheric pressure, temperature and precipitation go back 300 yr at most, and then in only a few localities in the Northern Hemisphere. In most areas, reasonably complete surface coverage exists for only the past 50–100 yr. Upper air soundings have been regularly made on a global basis for only the past two or three decades, and satellite observations are an even more recent innovation. Fortunately, information about climate before the beginning of instrumental observations can be obtained from study of a wide variety of physical and biological phenomena that respond in some way to climate. The "unique value of proxy data for studies of climatic change" has been recognised by the US Committee for the Global Atmospheric Research Program[1]. Among its principal recommendations is a programme for the preparation and maintenance of climatic data banks, and of analyses based on both current and palaeoclimatic data. It envisages the use of palaeoclimatic records for setting boundary conditions and for model verification and also views these records as our only documentation of the extreme states of which the Earth's climatic system is capable. The WMO–ICSU Joint Organising Committee[6] also states that "special efforts should be made to locate, make available, and analyze . . . a wide variety of human recordings and natural indicators of climatic conditions in earlier centuries and millennia . . .". Here, the emphasis is on the use of time-series analysis of proxy records to improve understanding of physical processes and to explore the possibilities of empirical prediction, as well as on verification of numerical models. According to the Geophysics Research Board[2], the design of more robust water-resource systems would be helped by palaeoclimatic research which "offers a promising approach to better estimates of past climatic variability and drought probabilities." Palaeoclimatic data thus promise to play an important part in further development of climate theory as well as in more immediate practical applications.

Because climate varies over a large range of time and space scales, it is important to identify the time periods and geographical areas in which data would be most critical in helping to understand the existing climate. The Earth is now, and has been for some 10,000 yr, in an interglacial climatic state, characterised by relatively high sea levels, relative warmth, and largely ice-free conditions in temperate latitudes. However, even the present interglacial can be subdivided into periods of distinctively different climate. The last globally significant change seems to have taken place about 3,000 yr ago. Widespread cooling and associated circulation changes at this time are fairly well documented in the Northern Hemisphere. Smaller but at least locally significant variations have occurred in more recent

Fig. 2 Rainfall estimation from tree-ring data. Total annual precipitation at Santiago, Chile, is calibrated with ring-width indices at a nearby site using simple linear regression methods (correlation coefficient 0.65). This relationship is first used to estimate rainfall for the calibration period of actual meteorological observations (above), and then for the total length of the tree-ring record (below). The dendroclimatic reconstruction compares fairly well with semiquantitative information from historical sources as given by Lamb[30].



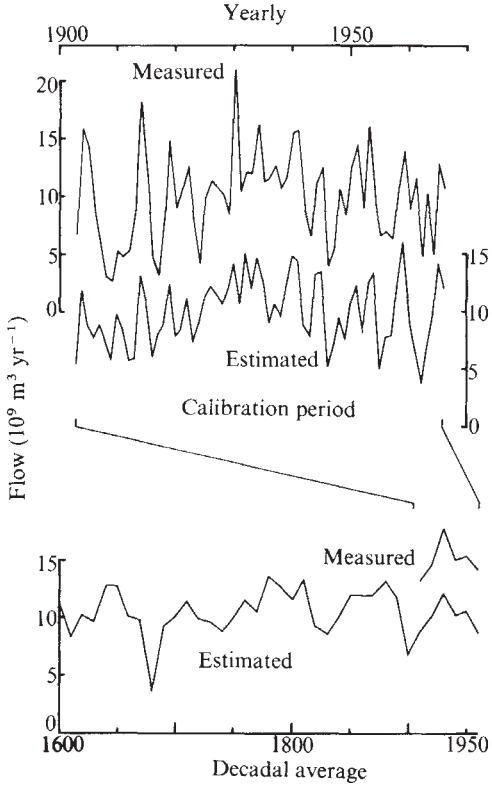
times. As pointed out by Kutzbach[7], palaeoclimatic records of the past several thousand years thus seem most relevant to study of the natural variability of our present climate, and fall within the scope of current planning. The kinds of natural variations we can expect in the near future will very probably be similar to those experienced in the relatively recent past.

Tree rings as climatic proxy data

Within this time framework, the annual rings of trees are one of the most valuable sources of climatic proxy data. Trees are found in most land areas of the Earth, they can reach ages of several hundred to a few thousand years, and under favourable circumstances their growth rings can be dated to the exact year of formation. Dendrochronology can provide long, continuous sequences of accurately dated annual rings, and is a dating tool that has found applications in fields ranging from archaeology to vulcanology. More important here is the potential for detailed palaeoclimatic reconstructions, because the physical and chemical properties of the annual rings can yield records of a broad range of environmental variables that are directly or indirectly linked to climate. Tree rings combine environmental 'sensitivity' and great length of record with the potential for resolution to individual calendar years, or indeed, to individual seasons. Thus, tree-ring-based proxy data series bridge the critical gap between the longer but more generalised geological records of climate, and the shorter and more detailed records of instrumental observations of climatic, geophysical, geochemical and astronomical variables that are critical to understanding the causes of climatic variability and change.

Dendroclimatology, concerned with the reconstruction of past climates through study of the properties of tree rings, has developed rapidly within the past 15 yr. Major factors have been the greatly increased speed of data processing possible using modern computers together with the development and application of powerful multivariate statistical techniques, greater understanding of the biological processes linking climate and tree growth, and the expansion of applications to broad new geographical areas. Exciting results are beginning to emerge from the use of a variety of new methods now being developed for investigation of the physical and chemical properties of the wood itself, such as stable isotopic composition[8-14], density[15,16] and amino acid composition[17]. However, although a large body of understanding and expertise has evolved for dealing with the climatic interpretation of ring-width variability, more recent techniques based on physicochemical properties of tree rings are still in their infancy. Moreover, study of the traditional dendroclimatic variable—the width of the annual ring—has yielded a large body of palaeoclimatic data that serves to illustrate the range of possible applications, the nature of long climatic proxy series, and the implications of past climatic variability for the future.

Fig. 3 Streamflow reconstruction. Tree-ring records from semi-arid sites in the Neuquén River basin in Argentina were incorporated in a canonical regression analysis[35] to estimate annual runoff at Paso de los Indios for the calibration period of gauging station record (above) and for the entire period 1600–1967 (below). Trees and drainage basins integrate fluctuations in temperature and precipitation in similar fashion, as shown by the good correspondence of decadal average values, and a correlation coefficient of 0.73 between measured and estimated annual values.



Ring-width variability

Many of the biological insights, recent technical developments, and palaeoclimatic applications of ring-width data have come through the work of H. C. Fritts and his associates at this laboratory. His book[18] gives a thorough review of the field, develops a methodology, and illustrates its use, primarily through examples from western North America. A good, botanically orientated review of tree rings as a natural data storage system has been published[19] in Great Britain with a lengthy bibliography, and the potential for application of dendroclimatology there is illustrated by recent findings[20-22]. There is also renewed interest in climatic and other applications of tree rings elsewhere in Europe[23,24] and in eastern North America[25], and a major effort is underway in the temperate latitude Southern Hemisphere[26], including Australia[27] and New Zealand[28].

The basic premise underlying the interpretation of ring-width variations in climatic terms is that climatic factors can directly or indirectly limit growth processes in trees[18]. The most suitable trees (with climatically 'sensitive' ring series) are often found in marginal situations, such as at the northern treeline, at the upper treeline in the mountains of temperate latitudes, near the arid forest border in desert regions, or as suppressed understory trees in temperate rain forests. Under such circumstances, in an adverse year, one or more environmental factors, such as soil moisture content, air temperature, or sunlight may frequently limit biological processes such as photosynthesis that are essential for development of an annual ring of normal width. Because all the trees in the same area experience the same general weather conditions, all those in appropriate ecological situations will form a narrower-than-usual ring in a bad year. The resulting synchronous sequences of wide and narrow rings provide the basis for cross-dating of tree-ring series (Fig. 1) as well as for palaeoclimatic inferences.

Data collection

A variety of approaches can be followed in extracting climatic proxy information from tree rings. A general procedure is first to select field collection areas and to identify suitable species, and then obtain tree-ring samples in the form of increment cores and cross-sections. Experience indicates that samples for ring-width analysis should include cores from 10 to 30 trees on each geographically limited and ecologically homogeneous site. After sample preparation, cross-dating of the tree-ring series (Fig. 1) leads to development of a site chronology, ensuring accurate dating of the annual rings. Measurement of ring-widths of selected samples follows. Transformations are then applied to individual radial series of ring-width measurements to remove biological trends. The next step is internal evaluation of the dendroclimatic potential of the site, based on the degree of common variation of tree-ring properties through time. Results of cross-correlation tests and analysis of variance, for example, indicate the degree of correlation between radii on the same tree and between trees on the same site, and are measures of the

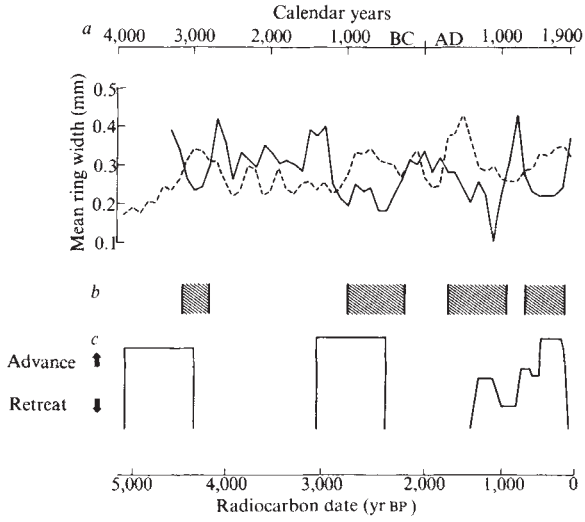


Fig. 4 Long-term fluctuations in temperature and precipitation. Although not explicitly calibrated in terms of climatic parameters, records from bristlecone pine in the White Mountains, California, offer strong evidence for changes in climate spanning centuries to millennia. (*a*), Successive 100-yr means of the average annual ring-width in large samples of trees from two ecologically distinct habitats, with quite different ring growth responses to climate. ——, Upper treeline (temperature-sensitive); ·····, lower forest border (drought-sensitive). Inferred periods of relatively cool and moist climate in the White Mountains (*b*) agree well with evidence (*c*) of periods of growth and extension of mountain glaciers, mainly in western North America and northern Europe. Preliminary data for the lower forest border from C. W. Ferguson (personal communication); upper treeline curve from ref. 36; Holocene glacial fluctuations from ref. 37.

common climatic signal present in the data. Finally, a mean value function is computed by averaging annual values of ring-width indices to produce the final index chronology for the site.

The kind and amount of climatic 'information' contained in a tree-ring data series for a particular site can be established by comparison with nearby weather records. Multivariate techniques have been used to develop a transfer function (response function) relating departures of total weekly, monthly, or seasonal precipitation and mean temperature to departures in the tree-ring variable of interest[29]. The weather data may be from individual station records or may represent regional averages. The nature of the response function will, to some extent, determine the kinds of palaeoclimatic reconstruction that can be attempted using a particular tree-ring data series or set of series. It should be emphasised, however, that a diversity in response functions for tree-ring data from a given region is often of real advantage for palaeoclimatic inference. This kind of diversity stems from differences in response between species, and between trees of the same species in different habitats. Thus, an effort is made to maximise the information contained in a regional tree-ring data array by sampling trees on sites representing a range of altitudes and microclimatic situations, and to develop tree-ring chronologies for trees of different species in the same areas. Possible differences in the trees' response, both to different climatic variables and to the same variables in different seasons of the year, can be tested by multivariate response function techniques. The results help in determining the kind of approach to be used in palaeoclimatic reconstruction.

Palaeoclimatic reconstruction

Several kinds of palaeoclimatic indices can be derived from tree-ring data. The kinds of climatic variables analysed, the spatial scale of the analysis, the degree of resolution in space and time, and the accuracy of the reconstruction or estimation will depend mainly on the inherent properties of the tree rings available for study and on the location of the trees. The proximity, quality and length of the available climatic records may also place some constraints on the scope of analysis. The quality of the proxy climatic records obtained from tree-ring data is probably best illustrated by examples of recent (and in some cases, preliminary) findings that range from estimates of a single variable at one point, based on a single tree-ring chronology, to the reconstruction of large-scale spatial anomaly patterns of several climatic parameters derived from analysis of data from a large network of tree-ring stations.

The simplest approach to palaeoclimatic reconstruction with tree-ring data is to use a highly stratified sample from trees responding (to a first approximation) in a simple manner to a single dominant climatic variable, such as precipitation. In this way, one can estimate past precipitation amounts at a point, along a meridional or zonal transect, or over a region. This kind of application is illustrated by a series of estimated annual precipitation amounts since AD 1010 for Santiago, Chile (Fig. 2). Inspection of this proxy record suggests that long periods of above- or below-average rainfall have occurred in the historically recent past, but are not well represented in the brief observational record. The ultimate extension of the single-variable approach is to use a regional array of tree-ring data points to infer past spatial anomalies in some dominant climatic parameter. This has been done successfully for precipitation[31,32] and for temperature[33] in the western US.

Complex, as well as simple, climatic variables can be reconstructed from tree-ring data. For example, stream runoff is often found to be positively related to precipitation, but negatively related to temperature, because warmer-than-usual conditions increase evapotranspirative water losses within the drainage basin. High temperatures can also limit growth processes in drought-sensitive trees directly as well as through depletion of soil moisture. Thus, tree-ring records can yield good proxy records of streamflow[34] as illustrated in Fig. 3, which shows large long-term fluctuations in flow from a major river basin in Argentina.

Past records of two variables can also be inferred at a single point, yielding information of great value for synoptic scale diagnostic analysis[36]. This is illustrated by comparison of very long records from bristlecone pine (*Pinus longaeva*) in the White Mountains of California (Fig. 4). In their lower altitudinal range,

Fig. 5 Biological impact of climatic variability. Bristlecone pines at upper treeline in the White Mountains of California are sensitive to fluctuations in warm-season temperatures. In a cool period (AD 1300–1850) coincident with Europe's historic 'Little Ice Age', the average growth rate dropped to half its former value (Fig. 4) and many trees succumbed to adverse growing conditions[38], leaving the trunks littering the foreground.

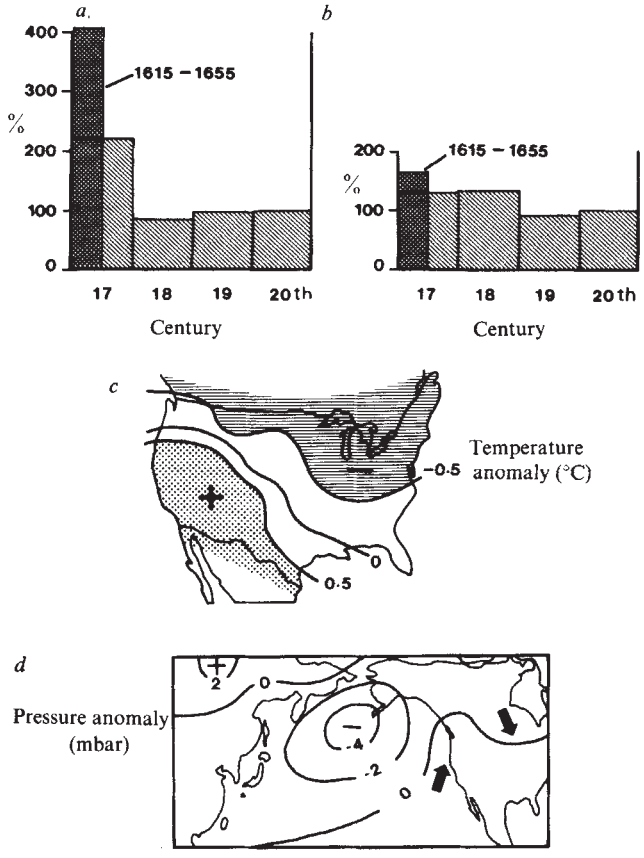


Fig. 6 Past climatic anomalies. Data from a grid of tree-ring sites in western North America have been used to reconstruct surface pressure anomaly patterns over the subhemisphere centred on the North Pacific Ocean for each season of each year, 1602–1962. Correlation analysis permits identification of the past occurrence of winters with pressure anomaly patterns like those of 1976–77 (*a*) and 1977–78 (*b*) over the past four centuries. Their apparent frequencies (above) are expressed as a percentage of their frequency in the twentieth century. Winters like these were especially common in the period 1615–1655. Reconstructed temperature (*c*) and pressure anomaly patterns (*d*) (based on 1901–1970 normal period) are shown below. Arrows indicate anomalous flow directions, bringing cold air to eastern North America. Adapted from Fritts and Lofgren[40].

ring-width is limited by drought conditions, especially low precipitation, whereas at upper treeline (Fig. 5) growth is mainly limited by low temperatures. Therefore, it is possible, by taking advantage of this diversity in climatic responses, to more fully characterise the nature of past climatic variations in this area. The warmth and humidity of the past century seem atypical in the light of the longer record.

Although ring-growth in trees can sometimes be approximated as a function of a single dominant climatic variable, a complex response to several environmental variables that are linked to climate is the more general case. Furthermore, the response to each variable, such as temperature, may change in magnitude or sign depending on the season of the year. For example, higher-than-usual temperatures in the previous autumn may be related to formation of a wider-than-usual ring in the current growing season, whereas high temperatures in the period immediately before and concurrent with the growing season may favour formation of a relatively narrow ring. Such relationships have been determined, and are readily interpretable in terms of the biological processes governing ring-growth[18]. Although comparatively little is yet known about the detailed relationships between climate and densitometric and isotopic variables, similar complex relationships are likely to be found in many cases[14].

Multivariate statistical methods have been successfuly applied to the simultaneous estimation or reconstruction of large scale spatial anomaly patterns of individual climatic variables, using tree-ring data representing a diversity of complex climatic response functions[29]. Techniques such as principal component analysis, stepwise multiple linear regression, and canonical regression analysis are used as tools for palaeoclimatic reconstruction on local to subhemispheric scales. Fritts *et al.* have been successful in using tree-ring data from western North America to estimate past sea-level pressure anomalies over a large area of the Northern Hemisphere[39]. In recent work, they evaluated the unusually cold (in the eastern US) winters of 1976–77 and 1977–78 in the perspective of several hundred years of proxy climatic record provided by tree rings. Their findings, shown in part in Fig. 6, strongly indicate that severe winters similar to these have occurred with great frequency at certain times in the relatively recent past. During part of Europe's historic 'Little Ice Age', the period 1615–1655, 70% of the winters seem to have been characterised by circulation patterns and associated temperatures over North America similar to those of one or the other of the past two winters.

Future directions

The future directions of dendroclimatic research will undoubtedly be strongly influenced by the needs of the climatological research community for specific kinds of proxy data in particular geographical areas. Nonetheless, present trends should lead to further expansion of the geographical data base, increasing contributions of new kinds of palaeoclimatic information from physicochemical analyses, and more intensive investigation of the time-series properties of tree-ring and derived data. The need to improve our ability to anticipate or predict future climate may result in the greater use of dendroclimatic data for actuarial-type analyses of the probability of future recurrence of past climatic episodes, and perhaps even for empirical forecasting.

This research was supported by the Climate Dynamics Program, Division of Atmospheric Sciences, USNSF.

1. US Committee for the Global Atmospheric Research Program *Understanding Climatic Change* (NAS, Washington, 1975).
2. Panel on Water and Climate *Climate, Climatic Change, and Water Supply* (NAS, Washington, 1977).
3. US natn. Acad. Sci. *Climate and Food: Climatic Fluctuation and US Agricultural Production* (Washington, 1976).
4. *Report of a Committee on Climatic Change* (AAS Rep. No. 21, Canberra, 1976).
5. *Living with Climatic Change, Phase II* (Science Council of Canada, Ottawa, 1976).
6. WMO/ICSU, GARP Organizing Committee *The Physical Basis of Climate and Climate Modelling* (GARP Publ. Ser. 16, WMO, Geneva, 1975).
7. Kutzbach, J. E. in *The Physical Basis of Climate and Climate Modelling* Appendix 1.3 (WMO/ICSU Joint Organizing Committee, GARP Pub. Ser. 16, WMO, Geneva, 1975).
8. Libby, L. M. *et al. Nature* **261**, 284–288 (1976).
9. Epstein, S. & Yapp, C. J. *Earth planet. Sci. Lett.* **30**, 252–261 (1976).
10. Pearman, G. I., Francey, R. J. & Fraser, P. J. B. *Nature* **260**, 771–773 (1976).
11. Gray, J. & Thompson, P. *Nature* **262**, 481–482 (1976).
12. Wilson, A. T. & Grinstead, M. J. *Nature* **265**, 133–135 (1977).
13. Lerman, J. C. & Long, A. *short Papers of the 4th Int. Conf. on Geochronology, Cosmochronology and Isotope Geology* (ed. Zartman, R. E.) 78–101 (US Geol. Survey Open-File Rep., 1978).
14. Long, A., Lerman, J. C. & Ferhi, A. *Short Papers of the 4th Int. Conf. on Geochronology, Cosmochronology and Isotope Geology* (ed. Zartman, R. E.) 78–701 (US Geol. Survey Open-File Rep., 1978).
15. Polge, H. & Nicholls, J. W. P. *Wood Sci.* **5**, 51–59 (1972).
16. Parker, M. L. *Syesis* **9**, 163–172 (1976).
17. Engel, M. H., Zumberge, J. E. & Nagy, B. *Analyt. Biochem.* **82**, 415–422 (1977).
18. Fritts, H. C. *Tree Rings and Climate* (Academic, London, 1976).
19. Creber, G. T. *Biol. Rev.* **52**, 349–383 (1978).
20. Hughes, M. *New Scient.* **77**, 500–502 (1978).
21. Hughes, M. K., Gray, B., Pilcher, J., Baillie, M. & Leggett, P. *Nature* **272**, 605–606 (1978).
22. Brett, D. W. *Weather* **33**, 87–94 (1978).
23. Eckstein, D. *Tree-Ring Bull.* **32**, 1–18 (1972).
24. Munaut, A. V. *Lejeunia* **91**, 1–47 (1978).
25. Cook, E. R. & Jacoby, G. C. *Science* **198**, 399–401 (1977).
26. LaMarche, V. C. Jr *Proc. WMO/IAMAP Symp. Long-Term Clim. Fluct.* 21–30 (WMO, Geneva, 1975).
27. Ogden, J. *Aust. J. Ecol.* (in the press).
28. Dunwiddie, P. *N. Z. J. Bot.* (in the press).
29. Fritts, H. C., Blasing, T. J., Hayden, B. P. & Kutzbach, J. E. *J. appl. Met.* **10**, 845–864 (1971).
30. Lamb, H. H.in *Polar Meteorology*, WMO Tech. Note 87, 428–453 (WMO, Geneva, 1967).
31. Fritts, H. C. *Mon. Weath. Rev.* **93**, 421–443 (1965).
32. LaMarche, V. C. Jr & Fritts, H. C. *Mon. Weath. Rev.* **99**, 138–142 (1971).
33. LaMarche, V. C. Jr & Stockton, C. W. *Tree-Ring Bull.* **34**, 21–45 (1974).
34. Stockton, C. W. *Long-Term Streamflow Records Reconstructed From Tree Rings*, Pap. Lab. Tree-Ring Research No. 5 (University of Arizona Press, Tucson, 1975).
35. Holmes, R. L., Stockton, C. W. & LaMarche, V. C. Jr *Water Res. Bull.* (in the press).
36. LaMarche, V. C. Jr *Science* **183**, 1043–1048 (1974).
37. Denton, G. H. & Karlén, W. *Quatern. Res.* **3**, 155-205 (1973).
38. LaMarche, V. C. Jr *Quatern. Res.* **3**, 632–660 (1973).
39. Blasing, T. J. & Fritts, H. C. *Quatern. Res.* **6**, 563–580 (1976).
40. Fritts, H. C. & Lofgren, G. R. *Quatern. Res.* (submitted).

Mechanisms and models of climatic change

G. J. Shutts & J. S. A. Green
Department of Physics, Imperial College, Exhibition Road, London SW7, UK

Many theories about climate change are essentially untestable, but we can still develop a consistent model based on understandable physics—in fact the data cannot be interpreted without one. Shutts and Green believe that there is some fundamental defect in all present models: it could be something physically improbable, like an unreasonable effect of tiny solar variations. More likely there is a fundamental lack in our appreciation of how very interactive systems behave.

CLIMATE modellers, in attempting to understand climatic an infinite number of mathematical models of physical processes and feedback mechanisms which are believed to be important. Economy of representation is achieved by the parameterisation of many physical processes. For instance, nearly all current models have avoided solving the equations for radiative energy transfer through the atmosphere and have made do with simple empirical schemes. Even though the theory of radiative transfer in the atmosphere is well known, its computation is tedious and parameterisation is simple and effective. Such an approach assumed that an assemblage of empirical processes (each one realistic in isolation) will behave like the real atmosphere, which has an infinite number of degrees of freedom and connections between processes. The art of parameterisation requires a blending of theoretical understanding, observation and physical intuition to achieve this and is a necessary part of modelling most, but especially very complicated, physical systems.

Definitions of climate

Is any useful purpose served by trying to invent a universal definition of climate? From a philosophical point of view, a definition in physics has little value outside the context of a mathematical theory. For example, the concept of heat becomes hopelessly ambiguous and devoid of use without the first law of thermodynamics, even though it retains its intuitive association with fire and physical sensation. Synoptic meteorologists go to some lengths to define 'blocking': a condition observed on maps of pressure patterns intimately related to anomalous weather conditions. Such people are unanimous in their diagnosis of a 'blocking situation' but are hard pressed to agree on a definition.

Over 20,000 years ago, nearly half the Earth was covered with ice to a depth of 1 km. There was a time a few hundred years ago when 'Frost Fairs' used to be held on the Thames River in London, and North West Europe suffered a drought between 1975–76. This is sufficient information to convey the subjective idea of climate change.

Climate is conventionally regarded as the state of the atmosphere averaged over an appropriate interval of time. If we believe that the climate changes in response to changes in 'external' processes, then the time interval can be chosen so that they vary little over its duration. By external processes we would certainly include variations of the intensity of solar radiation, the orography and perhaps also the state of the deep ocean and cryosphere. These definitions imply that the atmosphere is in a state of near-statistical equilibrium and responds to slow changes in the external factors.

The concept of quasi-static changes is a familiar feature of thermodynamics, and it is not surprising that many of the methods of statistical physics are 'borrowed' by the climatologist. Climate is sometimes thought of as an average over a hypothetical ensemble of atmospheric states that are in equilibrium with the slowly changing external factors. But, for instance, if the atmosphere always existed in one of two radically different states (glacial and non-glacial perhaps) this definition would hardly be appropriate. It would be much more useful to describe such a climate as the probability of the two states.

But what if climate changes are completely 'internal' and occur with fixed solar radiation, orography and ocean state (that is, are merely internal 'thermodynamic' fluctuations)? Choice of an averaging time interval is then more arbitrary, but at least should be sufficiently long to filter out 'weather' fluctuations. Definition of climate and its subdivision from medium-range weather should be left to the philosophers.

Hypotheses of climate change

Before launching into an account of the possible causes of climatic change, we will briefly describe the principle aspects of the general circulation of the atmosphere and its cause.

The atmosphere behaves as a heat engine in response to its nonuniform absorption of energy radiated from the Sun. The resulting spatial variation of internal and gravitational potential energy is converted into kinetic energy of air motion by processes such as ordinary convection. The distribution and intensity of solar radiation is controlled primarily by the geometry of the Earth and its orbit. Much of the incident radiation (about 30%) is reflected back to space by clouds, ice and the type of land surface. The atmosphere is quite transparent to this short wavelength solar radiation, with most of the energy occurring in the visible part of the spectrum. Energy balance is achieved by terrestrial re-radiation (at infrared wavelengths) to space from the uppermost 'optically black' layers of water vapour and carbon dioxide. The outgoing infrared radiation is of much smaller spatial variation than the incoming solar beam with the result that low latitudes receive a surplus of energy and middle and high latitudes a deficit. Balance of the integrated infrared radiation with the incoming solar (depleted by reflection mainly from clouds) determines the overall average temperature of the atmosphere—presently about 250 K. Vertical convection in the form of clouds transfers energy upwards from the ground to the high troposphere (about 12 km) and determines the vertical structure of temperature. It remains to transfer heat from equatorial to polar regions, and it is this process that determines the overall temperature variations at the surface. It turns out that the travelling depressions and anticyclones, characteristic of

Sorry, for copyright reasons some images on this page may not be available online

Frost Fair on River Thames, 1814

Mary Evans Picture Library

© Macmillan Journals Ltd 1978

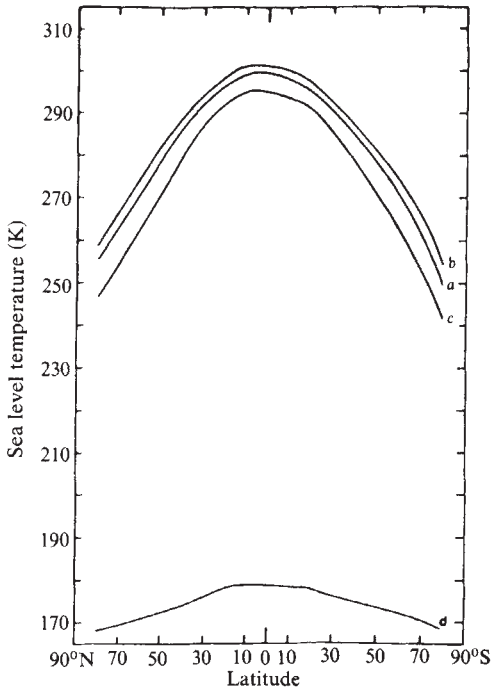


Fig. 1 Steady-state temperature distributions for various values of solar input and parameterisation[33] showing the calculated extreme sensitivity of surface temperature to small variations in solar input. Global mean temperature (K): *a*, 287.06; *b*, 289.81; *c*, 281.22; *d*, 174.65. Solar input: *a*, no change; *b*, +1%; *c*, −1%; *d*, <1.6%.

unsettled weather, are the primary agency for transporting heat polewards in the atmosphere. These large scale eddies are dominated by effects due to the Earth's rotation and tend to redistribute angular momentum, causing the predominance of surface westerlies in middle latitudes, and easterlies elsewhere. These wind belts in turn determine the climatic belts of the planet.

Mountain ranges and heat anomalies, like warm winter oceans, tend to obstruct the circulating air, thereby creating stationary wave motion. The resulting nearly steady motion accounts for much of the zonal asymmetry of the climate. The Asiatic winter anticyclone and summer monsoonal low pressure are two outstanding examples of this type of local climatic variation.

(1) Solar energy output changes: The most obvious hypothesis for climate change attributes fluctuations in the global mean temperature to variations in the energy output of the Sun. Observations only in the past few decades have barely been able to detect changes in the solar irradiance of the order of 0.1% over 27 days. Several investigators[1,2] have correlated periods of high solar activity (indicated by the presence of sunspots) with anomalous tropospheric weather events such as 'blocking' or thunderstorm frequency.

The infamous 11-year sunspot cycle has been found to be astonishingly well correlated with a variety of atmospheric parameters (for example, hot summers, rainfall, lightning frequency) but in the absence of a plausible physical theory to account for them, scepticism remains.

(2) Orbital changes affecting solar insolation: Palaeoclimatological records for the last million years (the Pleistocene period) and particularly estimates of the global ice volume changes by analysis of the isotopic oxygen content of fossil plankton, indicate a strong periodicity of 100,000 years. Milankovitch[3] proposed that the secular variations of the Earth's orbit about the Sun would be sufficient to give fluctuations in the solar constant of a few per cent.

These orbital changes are due to changes in (i) the eccentricity of the orbit with a period of 96,000 yr; (ii) the obliquity of the Earth's axis with a period of 40,000 yr; (iii) the Earth's tilt by precession, with a period of 26,000 yr; and (iv) the orientation of the elliptical orbit which rotates with a period of 96,000 yr.

Mason[4] agrees that the period and magnitude of these orbital changes lead to solar energy fluctuations which can explain parameters of the palaeoclimatological record, such as global ice volume change.

(3) Greenhouse effect—modification by CO_2: Global increases in the amount of CO_2 in the atmosphere through the burning of fossil fuels are thought to be important enough to influence the greenhouse effect (the process by which infrared radiation is trapped by CO_2 and water vapour leading to surface temperatures higher than the black body temperature). Although the direct effect of an increase in CO_2 may be quite small (surface temperature rises of about 2 K if the amount of CO_2 is doubled), it is difficult to estimate how the rest of the system will react. For instance, a small global temperature increase with the addition of CO_2 will increase the moisture holding capacity of the air. Water vapour is the most important constituent involved in the greenhouse effect and so the net increase in temperature could be much greater than that predicted on the basis of CO_2 alone.

(4) Volcanic dust: The presence of volcanic dust in the upper atmosphere has been put forward as a way of increasing the albedo and thus reducing the mean atmospheric temperatures. Hunt[5] has investigated such influences with a three-dimensional general circulation model and has found a mean surface temperature reduction of at most 1 K. He concludes that volcanic eruptions could have a significant effect on the atmosphere, though lasting changes are improbable.

(5) Ice cap/albedo positive feedback: Climatology literature abounds with possible positive feedback mechanisms which could render the climate unstable. Of these, the polar ice cap/albedo feedback is the most popular. The argument is quite simple. An increase in the horizontal extent of the polar ice sheet increases the global albedo, which in turn results in lower temperatures through the reduction in absorbed solar radiation. Colder surface conditions encourage the expansion of the ice sheet and so on. Budyko[6] and Sellers[7] have investigated the equilibrium mathematical solutions of simple zonally averaged heat balance models which have ice sheet dependent albedo variation and include a diffusive poleward transfer of heat. Their calculations show that a decrease in the solar irradiance of 2–5% would be sufficient to initiate another ice age.

(6) Oceanic influences: Weyl[8] has presented a case for the initiation of Ice Ages by reduction of the salinity of the North Atlantic Ocean. He claims that the 'Little Ice Age' of AD 1500–1850 substantiates his hypothesis. Certainly on a time scale of years or tens of years the ocean surface temperatures are believed to exert a strong influence on weather types. Namias[9] believes that the evolution of monthly mean weather is largely controlled by the oceanic temperature anomalies.

(7) Other topographical influences: Changes in the condition of the Earth's land surface by mountain building and vegetation growth or decay are amongst other possible climate change mechanisms. Charney[10] has described a positive feedback mechanism which causes desert conditions through overgrazing: reduction of plant cover increases the albedo which, in turn, causes a greater radiative deficit. The resulting horizontal temperature gradients induce subsidence which inhibits convective precipitation and therefore curbs plant growth.

Methods of climate modelling

It is possible to divide climate models into three broad categories: (1) energy balance models that include radiation/albedo feedback processes but contain no explicit dynamics; (2) low resolution statistical–dynamical models that include large scale eddy transports of heat and momentum in parametric form, for example, as some curious variety of conductivity; and (3) simulation models of very high resolution that include all motion on scales greater than some 100–200 km.

(1) The energy balance models can be used to explore the sensitivity of zonally averaged climatic states that are determined by the ice-cover/albedo feedback mechanism (see Fig. 1).

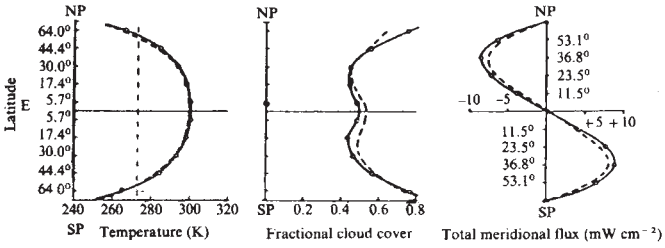


Fig. 2 Taken from Paltridge[12]. ——, Predicted; ----, observed.

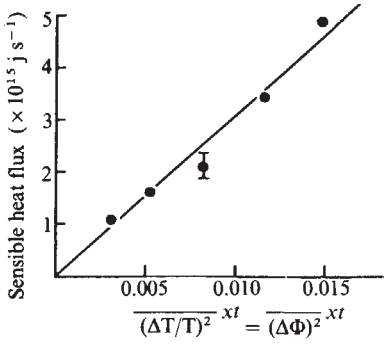
They show that small changes in the solar input can cause catastrophic changes in the surface temperature of the model. Using a time-dependent model, Schneider and Gal-Chen[11] have found that a reduction in the intensity of solar radiation by more than 1.6% results in a fall of global mean temperature of over 100 K. As with all climate models, it is difficult to believe that such large excursions from present conditions are adequately represented by the processes, the parameterisation of which depends on the current state of the atmosphere.

A unique approach was adopted by Paltridge[12] who sought a minimal principle governing the zonally averaged distribution of cloud cover, surface temperature and poleward heat transport. The incident solar radiation at the top of the atmosphere was the only externally prescribed quantity. By minimising the rate of generation of entropy, very close agreement of the fundamental model unknowns with observation was found (Fig. 2). No dynamical information was required in the model—a disturbing fact when so much effort is devoted to integrating the equations of fluid motion. Further work by Paltridge[13] has led to predictions of the global cloud cover and rainfall, again with surprising accuracy.

The notion that the climate does not depend on the Earth's rotation seems hard to accept, yet the agreement with observation is striking. The philosophy of such an approach is very appealing, and is in line with the idea of applying thermodynamic concepts to the climatic problem. Indeed, the principle of minimum rate of generation of entropy is well established in the field of non-equilibrium thermodynamics for a certain class of physical processes[14].

(2) These models are primarily concerned with the main global features of the time-averaged wind and temperature fields without explicitly describing the transient depressions and anticyclones or any other smaller scale of motion. The statistical effect of these smaller scales of motion is parameterised in terms of the larger scale fields which are described explicitly. One such 'prescription' for the turbulent transfers of heat and momentum accomplished by large scale eddies was given by Green[15]. Some aspects of this parameterisation seem to describe global-scale transports realistically (Fig. 3). The parametric formulation of the eddy transport of angular momentum is not so realistic and remains one of the most important and difficult aspects of our understanding which is highlighted in these models.

Fig. 3 Eddy flux of sensible heat across 50°N plotted against $\overline{(\Delta T/T)^2}^{xt}$; $(\Delta T/T)$ = fractional temperature difference between 20 and 70°N, 500 mb (from Green[15]).



(3) The general circulation models of this category are widely used for climatic studies. They are essentially similar to the models used for weather prediction and seek to embrace as many relevant physical processes with as much resolution as is computationally feasible. In contrast to the previous type, these models extend their explicit representation to that of weather patterns, even though in their weather forecasting capacity they have predictive value for only a few days. In the most advanced models, the circulation of the oceans is also modelled, though integrated with a much longer time step to that of the atmospheric circulation.

Uses and abuses of models

The Budyko–Sellers models are useful for assessment of sensitivity: most of the estimates of the effect that changes in CO_2 or solar irradiance might have are based on them. Unfortunately, there is only the insufficient climate record against which they might be tested. This is true of most models and has the corollary: "any fool can make climate predictions because he won't be there when it's time to collect the bad debts."

The Paltridge model apparently gives excellent results, but no-one understands the physics that goes into its construction. Furthermore, it has only been tested against the present climate of the Earth, and might, therefore, be just a one-off success. Highly parameterised models can reproduce some results of more complex models (though not as well as that of Paltridge) but so far are generally considered to be an outside chance in view of their palpable over-simplification. General circulation models (GCMs) have been extensively used. (The UK Meteorological Office spends about as much on running their model as goes into university education of meteorologists in Britain.) Much is known about their behaviour which is not known about other models. The nature of these characteristics may form a guide to those of other models.

All models are much more similar to each other than they are to the real atmosphere—it's easier to forecast the weather forecast than the weather. Details are consistently wrong in all models.

Although the predicted values for the poleward transport of energy and momentum are surprisingly accurate (they are closer to the long-term mean than the year-to-year variability of atmospheric transfers), the calculated intensity of the eddy motion is much less than observed.

Earlier versions demanded some judicious manipulation of parameters in order to reproduce the present climate. That this is said not to be true of later models may indicate greater skill in anticipating the required numbers or greater understanding of the relevant physical processes. How can we tell?

All GCMs produce westerly jet streams that extend much too far into the upper atmosphere. A detail, but it goes with not having the polar stratosphere warm enough by some considerable amount. It is worrying that this type of defect is tangible and universal in a region where we 'know all the physics'.

No climatic model has actually predicted a climatic anomaly or has been able to simulate one for which there was sufficient meteorological data to make it a proposition in numerical analysis. We believe that this applies, for example, to the anomalous North American winter of 1977–78, and to the British drought of 1975–76.

Finally, these models suggest that climate is essentially unpredictable: that we can only predict in the probability sense of many 'realisations' (or re-runs from 'nearly-the-same' initial data). Is it the climate model or climate that is unpredictable? We must stress that these conclusions are biased and arguable. Some are demonstrably true of general circulation models, but we believe many are symptoms of the climate problem.

1. Willet, H. C. *Geografiska Annlr.* **31**, 295–315 (1949).
2. Brooks, C. E. P. *Q. Jl R. meteor. Soc.* **60**, 153–165 (1972).
3. Milankovitch, M. *Dandbuch der Klimatologie I*, Teil A (ed. Koppen, K. & Geiger, R.) (Gebr. Borntraeger, Berlin, 1930).
4. Mason, B. J. *Q. Jl R. meteor. Soc.* **102**, 473–498 (1976).
5. Hunt, B. G. *Mon. Weath. Rev.* **105**, 247–260 (1976).
6. Budyko, M. I. *Tellus* **21**, 611–619 (1969).

7. Sellers, W. D. *J. appl. Meteor.* **8**, 392–400 (1969).
8. Weyl, P. K. Role of Oceans in Climate Change, *Meteor. Monogr.* **8**, No. 30, 37–62 (American Meteorological Society, Boston, Mass, 1968).
9. Namias, J. *J. appl. Meteor.* **11**, 1164–1174 (1972).
10. Charney, J. G. *Q. Jl R. meteor. Soc.* **101**, 193–202 (1975).
11. Schneider, S. M. & Gal-Chen, T. *J. geophys. Res.* **78**, 6182–6194 (1973).
12. Paltridge, G. W. *Q. Jl R. meteor. Soc.* **101**, 475–484 (1975).
13. Paltridge, G. W. in *Climate Change and Variability* (ed. Gubbin, J.) 249–255 (Cambridge University Press, 1978).
14. Prigogine, I. *Introduction to Thermodynamics of Irreversible Processes* (Charles C. Thomas, Springfield, Illinois, 1955).
15. Green, J. S. A. *Q. Jl R. meteor. Soc.* **96**, 157–185 (1970).

Numerical simulation of climate and climatic change

A. Gilchrist
Meteorological Office, London Road, Bracknell, Berks, UK

Detailed three-dimensional numerical models of the atmosphere, coupled as necessary to models of other parts of the climatic system, provide the most promising approach to understanding the physical basis of climate. Models of this kind can be used to investigate the impact of anthropogenic pollution on climate. At the present time, the main concern is with increasing concentrations of CO_2 which might lead to overall warming of the troposphere, but chemical and thermal pollution may also pose a threat. The possible climatic changes would take place slowly and would involve the response of the slowly reacting parts of the climatic system, particularly the oceans. The problem of how to simulate such changes of climate presents many difficulties, which are currently being studied.

THE most powerful tool available for studying the physical basis of climate and climatic change is the general circulation model, which simulates the evolution of the atmosphere numerically and is capable in principle of taking into account all the factors that determine climate. Such models are complex and require a range of expertise and prolonged experimentation to bring them to the point where they can be put to good use in climate experiments that are of value. Consequently, their development has been undertaken at only a few centres. At a recent conference in Washington (April, 1978) organised by the Joint Organising Committee for the Global Atmospheric Research Programme for the purpose of comparing climatic simulations, results were presented from nine institutions. Of these, four were from the USA, two from the Soviet Union, with one from Australia, Canada and the UK. However, some of the models described were at a preliminary stage of development and only the American and UK Meteorological Office groups have tested models extensively in studies of climate.

Fundamental physical equations

The starting point for general circulation models are the fundamental physical equations for the rates of change of the components of velocity, temperature and humidity at a point in the atmosphere. The equations include a description of the processes by which the atmosphere interacts with radiation (infrared and solar) and the underlying surface of the Earth. They are solved on a network of points over the globe to provide a continuing evolution of atmospheric states. Processes that take place on too small a scale in the real atmosphere to be represented explicitly are 'parameterised' in terms of the variables available on the mesh; for example, the vertical transfer of heat and moisture due to buoyancy forces which in the real atmosphere produce cumulus clouds and showers is treated in this way.

Conditions that are external to the atmospheric dynamical system are specified. These include the constituents of the atmosphere, the radiation conditions, defining either explicitly or implicitly the Sun's position with respect to the Earth, and the characteristics of the Earth's surface that influence its interaction with the atmosphere and radiation. For a land point, the latter would include the height above mean sea level, the albedo without snow cover and the change due to snowfall, the roughness, and a heat capacity to define how its temperature would respond to the absorption of heat.

For a sea point a purely atmospheric model lacking a dynamical interaction with the ocean requires either the sea-surface temperatures to be specified or to be determinable from the local heat balance. The former, with the values set to the long-term climatological average, is appropriate when the object is to provide a realistic simulation of climate, and the latter has been used in attempts to understand particular features of the atmosphere's general circulation where the use of temperatures that are themselves partly determined by the atmospheric circulation is inappropriate. Combined ocean–atmosphere models which allow a dynamical evolution in both fluids are now being constructed and tested[1], as for many processes of interest in climate an interactive system is essential.

General circulation models

These provide in principle, and already to some extent in practice, a method of estimating the global climate for any set of prescribed 'external' conditions. To do so, they have to be integrated through an extended period of simulated time during which many atmospheric states are created. The length of integration needed may well depend on the precise problem being considered. It should be long enough for a statistical equilibrium to be established between the atmosphere and the prescribed conditions and, ideally, it should include a sufficiently varied range of atmospheric states to make it possible to estimate the probability of occurrence of any particular state with these conditions.

Given this information, the more usual characteristics that are taken to define climate may then be derived; these are the long-term averages, the extremes, the year-to-year variability, and so on. For most problems involving a change of climate, it would probably be more convenient to estimate the climate, not in absolute terms, but with reference to what is now observed. This is an easier goal, and as minor deficiencies revealed by model estimates of the present climate can probably be allowed for, the end-results may be achieved with rather simpler models.

Models are still in a comparatively early stage of development and up to the present attention has for the most part been concentrated on their ability to simulate the average conditions. For example, a number of simulations of the present climate have been described in the literature[1,3–6], and there have been attempts to simulate the climate of the last ice-age by using data from the CLIMAP project[7] to define the extent of glaciation and the sea-surface temperatures 18,000 yr ago. In general terms the

0028–0836/78/0276—0342$01.00 © Macmillan Journals Ltd 1978

models achieve a simulation of present climate that contains all the major features of the atmospheric circulation, but there are systematic errors in the intensity of some of them that call for further development and improvement.

The importance of this aim has been recognised by the international community which has supported international experiments such as GATE (the GARP Atlantic Tropical Experiment) in 1974 and FGGE (the First GARP Global Experiment) in 1978–79, the main purpose of which has been to obtain the data required to test and systematically improve the simulations. An important aspect of the FGGE is that data will be collected from a complete yearly cycle, enabling the models to be tested in a wide variety of conditions. Remembering that within the annual cycle there is a greater range of climates than have affected any particular part of the year at least in the last millenium, it is obvious that the potential of the FGGE to improve climatic modelling, provided a sufficiently comprehensive coverage of observations is obtained, is substantial.

The ice-age simulations[8–10] were obtained using similar methods to those applied to present-day climatology. Maintaining the sea-surface temperatures, the distribution of ice and the solar parameters at their 18,000 yr BP values, the general circulation models were integrated long enough for the atmospheric motions to come into statistical equilibrium with the boundary conditions. There are significant differences in the simulations, and because of a number of inadequacies the results must be considered indicative rather than definitive. Nevertheless, they are of interest in suggesting certain aspects of the ice-age climate that can be checked against geological evidence. Thus, the GFDL integration[8] favours comparative dryness in the ice-age tropics and this is considered to be supported by evidence provided by the sand-dune distribution of 18,000 yr BP. The model integrations agree in having generally less rain and lower global temperatures during the ice-age than now.

Why numerical models?

It is evident that although general circulation models are a very powerful tool for investigating the physical basis of climate and climatic change, they involve a lot of computing, and therefore it is tempting to enquire whether there might not be a simpler method which could be equally successful. There are, however, good reasons why large numerical models occupy a central position in studies of climate. The first stems from the fact that controlled experiments cannot be carried out on the atmosphere, at least not on anything approaching a global scale. In these circumstances, large numerical models provide the only feasible method of testing hypotheses. Without them research in climate can easily become (and indeed for a long time was) a series of uncheckable speculations. This is not to imply that the results of a particular model are correct; but it is an inherent property that they provide not simply an initial cause and a subsequent effect, but a chain of consequences linking the two. Taken individually, the links in the chain are usually of such a kind that they can be checked quantitatively against atmospheric observations; in particular, critical links can be identified. Another reason is that it has become clear that climate is determined by complex interactions among a number of systems, these being the atmosphere, the oceans, glaciers, the Earth's surface, and so on. Indeed, the essence of the problem of climate which distinguishes it from, for example, problems in atmospheric dynamics, is that interactions between the atmosphere and other systems are involved. The processes occurring within the systems are themselves often exceedingly complex; to rationalise and explain the interactions between them is generally speaking beyond the expertise of any one individual and can involve so many factors whose importance cannot be ruled out *ab initio* that the only possible approach is to try to incorporate them in a model that can be solved numerically.

These points can be illustrated from the work of Charney[11] in trying to explain the severe drought in the Sahel in 1972. The range of possible interactions that might have created the drought is very large, and they cannot be subjected to precise experiments; it may therefore be very misleading to choose certain of them as dominant without quantitative modelling of all the potentially important factors. In the first instance Charney[11] sought to explain the Sahel drought by a biogeophysical feedback mechanism. He postulated that a reduction of vegetation in a borderline desert region had occurred, possibly due to lack of rain or to heavy cropping by animals. A greater amount of light soil would therefore be exposed and the increase in the albedo would lead to a greater proportion of the incoming solar radiation being reflected back to space. Less heat would therefore be available to heat the atmosphere. If temperatures further from the desert remained unchanged, a more intense circulation with downward motion over the area of increased albedo would tend to be produced, and so reduce the rainfall and the vegetation cover even further. Charney proposed that this could account for the prolonged drought in the Sahel.

The hypothesis was later tested in experiments with a general circulation model[12]. In this case the situation was shown to be much more complex than the original explanation allowed for. Thus, it was found that, as expected, a larger albedo reduced the proportion of the incoming solar radiation absorbed at the surface, but because it also reduced the input of heat and moisture to the lower atmosphere there was less cloud. More solar radiation therefore, reached the ground, and in some experiments the absorption of solar radiation at the Earth's surface actually increased. However, with reduced cloudiness there was less infrared radiation back to the surface, with the result that the net absorption of radiation was less. The proportions that went into evaporation and into sensible heating of the atmosphere depended on the wetness of the surface; in the experiments this factor was fixed from climatological values, but in reality it depends on the rainfall, thus introducing another feedback mechanism that is potentially very important. In general, increasing the albedo led to a decrease in the precipitation, but this was not invariably the case. In experiments in which the evaporation was almost negligible and which therefore accorded most nearly with Charney's original conjectures, the results varied among the regions where the albedo was increased; the change seemed to increase precipitation in some areas and reduce it in others.

It is clear from Charney's experiments that the consequences of an increase in albedo are by no means easy to predict. An assessment of the physical processes involved requires rather precise quantitative modelling and a careful check against observations of all the interactions that seem to be important. From the point of view of explaining the Sahel drought the situation is even more complex, as there are factors other than the albedo that may be important. Thus, Walker and Rowntree[13] have shown that the soil moisture content can have a direct influence on atmospheric dynamical developments in desert fringe areas, and this factor was not represented effectively in Charney's experiments. Also, from studies of the effects of anomalous sea-surface temperatures on the atmosphere, carried out using the Meteorological Office general circulation model, there is some evidence that the droughts may be linked with abnormally cold water in the tropical Atlantic. If this is confirmed it might account for the longevity of some droughts in the Sahel, as there is considerable evidence of long period variations in tropical sea temperatures.

The processes that determine climate evidently involve delicate balances among physical systems which even in isolation are complex and not easily understood, and which cut across the boundaries between the individual expertise of scientists. The problems can only be solved as a result of effective cooperation between meteorologists, oceanographers, radiation experts, plant physiologists, and so on. General circulation models provide a means of integrating the contributions of each field of expertise to arrive at the total influence on climate.

Impact of man's activities

In considering climatic change due to natural causes, the existence of climatic data for the past, either as direct measurements

or as proxy data derived from tree-rings, ice-cores or cores from ocean or lake sediments, provides at least a statistical account of the variations that are likely. When one turns to the possible impact of man's activities, there is little chance of finding a sufficiently close historical analogy, and it is difficult to envisage a systematic approach to estimating the effects quantitatively which does not depend essentially on simulating them by means of sufficiently comprehensive numerical models. At the present time, for example, the main concern arises from the increase in CO_2 concentrations in the atmosphere, thought to be due principally to the burning of fossil fuels. One-dimensional column models, which calculate the local temperature change in a still atmosphere due to varying concentrations of the radiatively active gases, indicate that, for typical conditions, increasing the concentration of CO_2 produces a warming of the troposphere and a cooling of the stratosphere and upper atmosphere.

There is now substantial agreement[14] among such models that the temperature change to be expected from a doubling of the amount of CO_2 from its present value of about 320 p.p.m is about 2 K at the Earth's surface. This value includes a feedback from water vapour which accounts for about half the total increase. The CO_2 warms the lower atmosphere, which in consequence can hold more water vapour, and the absorption of radiation by the water vapour enhances the warming. The results depend on the conditions of cloudiness assumed, and in general no allowance is made for changes in cloudiness that might arise from the increased temperature and water vapour content.

A general circulation integration by Manabe and Wetherald[15] using a much simplified surface topography and ignoring the effects of ocean currents, the annual solar cycle and variations in cloudiness, has demonstrated additionally that in higher latitudes, where the warming reduces or removes highly reflecting snow-cover with its typical low level atmospheric inversion, the change in surface temperature could be much greater, perhaps up to about 10 K. They also found that the hydrological cycle was more intense because of the additional moisture-carrying capacity of the atmosphere, and it can be surmised from their paper that the distribution of rainfall was significantly altered. A statement approved by the Executive Committee of the World Meteorological Organisation[16] has called particular attention to the CO_2 problem, and has pointed out the urgency of defining its impact more precisely.

The immediate problems are, firstly, to forecast future concentrations of CO_2 and, secondly, to translate their effects into likely changes in climate. It is on the second problem that meteorological investigations must concentrate, although it has to be recognised that changes in climate may affect the storage of CO_2 in the oceans and in the biomass; therefore, the two aspects are not independent. The obvious first extension from present results is to try to obtain a clearer view of the regional changes, particularly of temperature and rainfall, that might be produced by increased amounts of CO_2. It is perhaps salutary to recall the results of a Meteorological Office experiment in which the Arctic was maintained ice-free[17]. This of course led in general to higher temperatures. Nevertheless, Britain and Western Europe were much colder because the predicted atmospheric flow patterns showed prevailing winds that were backed from their present direction towards a more easterly point. In a similar way, the effect of more CO_2 will probably vary from one part of the globe to another, some parts showing a cooling in spite of the global temperature being higher. There might also be significant changes in the probabilities of extreme events, such as the incidence of very cold winters or severe droughts in Britain, or of flooding of low-lying areas. It is important to try to identify these regional variations, with perhaps special emphasis on the rainfall changes which could easily have a substantial impact on agricultural productivity.

If present indications are correct, the effect of greater CO_2 concentrations will be felt increasingly over the next century or so. There will be a gradual change in the Earth's climate involving changes in the oceans, in sea-ice, in glaciers and in the vegetative cover of land surfaces as well as in the atmospheric motions. Just how radical the changes will be is uncertain at this stage. To produce a forecast of the evolution of the climate with time is a problem which may not be capable of solution in an exact sense, and for which indeed no clear methodology that would lead to a solution has yet been established. To create a model involving all the systems analogous to an atmospheric general circulation model and simulate their interactions comprehensively in a long-term integration, for example, is out of the question.

Estimates of future climate

There are, nevertheless, approaches that might be tried. It has already been emphasised that general circulation models provide a means of estimating the climate for a given set of 'external' conditions. We have to reckon with the fact, however, that few parts of the climatic system are truly external and independent of the atmosphere. Factors emanating from outside the Earth, such as solar radiation, factors emanating from below the Earth's surface, such as volcanic dust, and certain inputs by man himself, come into this category. There remains a complex of systems that interact with the atmosphere; some react quickly, whereas others change only slowly as a result of an atmospheric input.

The atmosphere itself is highly variable from day to day, and some aspects of the processes at the Earth's surface such as evaporation and the establishment or disappearance of temporary snow-cover are on a similar time scale. At the other end of the scale, glaciers and the deep ocean may require years of altered conditions to change significantly. From the point of view of establishing an estimate of climate that is essentially steady (apart from the annual cycle) over periods of years or decades, the very slowly responding systems can probably be considered as effectively external. A reasonable methodology for dealing with slow long sustained climatic change then emerges.

The first step would be to establish the statistical characteristics of a climate for a relevant set of 'external' conditions using a general circulation model that includes factors varying on short time scales. As well as the atmosphere and some land surface variations, this ought to include also the upper layers of the ocean, which react quickly to atmospheric changes. The existence of upwelling in the oceans by which in limited localities the deep waters can come fairly rapidly into contact with the atmosphere, might cause difficulties which would have to be overcome. The climate once established could then be allowed to act on the more slowly varying parts of the climatic system for as long a period as was deemed reasonable. For example, the gradual change in the ocean circulation and changes in glaciation could be calculated on the basis of a steady climate as long as the changes were not too large.

The assumption is essentially that for small changes the variations in the slowly responding systems are nearly linear. The extent to which this is so could of course be checked by preliminary experiments. Once significant changes had taken place it would be necessary to go back to the general circulation model to re-establish a relevant climatology. In this way it is at least methodologically possible to try to estimate the evolution of climate taking place over long periods of time.

The CO_2 problem is the most pressing at present, because there is enough generally accepted scientific evidence to indicate that the impact on climate could be substantial. Changes that might result from other human activities are less clearly understood, but there are obvious possibilities that they might become more important in the future. Thus, the input of waste heat into the atmosphere has significant effects locally, evident for example in the increase in temperature at some meteorological sites as a result of urbanisation nearby.

The future global energy position is far from clear, but if this involved the use of very large nuclear power stations, vast amounts of heat could be injected into the atmosphere in concentrated sources whose effects are as yet not well under-

stood. An experiment using the Meteorological Office general circulation model, carried out in association with the International Insitute for Applied Systems Analysis, has shown that the input of 1.5×10^{14} W to a particular area had a large effect on the global climate and changed the pattern of air movement and rainfall very substantially. This amount is much greater than present world consumption (~ 7–8×10^{12} W); nevertheless, the response is such as to demonstrate that it would be prudent to investigate the possible impact on climate before large energy projects which might release large amounts of energy are embarked on.

The possibility that the ozone layer might be reduced by the injection of pollutants into the stratosphere has been much discussed in recent years, first in the context of supersonic aircraft flying in the stratosphere and more recently as a result of the accumulation of inert gases such as freons. The possible danger is firstly to living material because of the increased amounts of ultraviolet radiation reaching the ground, and secondly, in the longer term, to possible changes in climate that this would produce.

Some of the early fears about increased radiation were evidently exaggerated[19], and the mechanism by which changes in ozone would lead to noticeable changes in the climate is far from clear. Some experiments[20] in which the ozone was increased showed substantial warming at the ozone levels themselves but little impact at the Earth's surface. Should the possibilities for climatic change be shown to be more real than has so far been demonstrated it would be necessary to follow a methodology similar to that just outlined for CO_2 to make some estimate of the detailed effects over a long period.

1. Manabe, S., Bryan, K. & Spelman, M. J. *J. phys. Oceanogr.* **5**, 3–29 (1975)
2. Holloway, J. L., Jr & Manabe, S. *Mon. Weath. Rev.* **99**, 335–370 (1971).
3. Stone, P. H., Chow, S. & Quirk, W. J. *Mon. Weath. Rev.* **105**, 170–194 (1977).
4. Washington, W. M., Otto-Bliesner, B. & Williamson, G. January and July simulation experiments with the 2.5° latitude–longitude version of the NCAR general model (NCAR/TN 123 STR, NCAR, Boulder, Colorado, 1977).
5. Gates, W. L. & Schlesinger, M. E. *J. atmos. Sci.* **34**, 36–76 (1977).
6. Corby, G. A., Gilchrist, A. & Rowntree, P. R. Methods in Computational Physics, **17**, *The UK Meteorological Office 5-level General Circulation Model.* (Academic, New York, 1977).
7. CLIMAP Project Members. *Science* **191**, 1131–1138 (1976).
8. Manabe, S. & Hahn, D. G. *J. geophys. Res.*, **82**, 3889–3911 (1977).
9. Williams, J., Barry, R. G. & Washington, W. M. *J. appl. Meteor.* **13**, 305–317 (1974).
10. Gates, W. L. *J. atmos. Sci.* **33**, 1844–1873 (1976).
11. Charney, J. G. *Q. Jl. R. meteor. Soc.* **101**, 193–202 (1975).
12. Charney, J. G., Quirk, W. J., Chow, Shu-Hsien & Kornfield, J. *J. atmos. Sci.* **34**, 1366–1385 (1977).
13. Walker, J. & Rowntree, P. R. *Q. Jl. R. meteor. Soc.* **103**, 29–46 (1977).
14. Augustsson, T. & Ramanathan, V. *J. atmos. Sci.* **34**, 448–451 (1977).
15. Manabe, S. & Wetherald, R. T. *J. atmos. Sci.* **32**, 3–15 (1975).
16. *WMO Bull.* No. 319 (1976).
17. Newson, R. L. *Nature* **241**, 39–40 (1973).
18. Williams, J., Kromer, G. & Gilchrist, A. *IIASA Research Memorandum* RM-77-15, 2361, Laxenburg, Austria (1977).
19. Groves, K. S., Mattingly, S. R. & Tuck, A. F. *Nature* **273**, 711–715 (1978).
20. Mattingly, S. R. Met. Office 20. Tech. Note II/123, Bracknell, UK (1978).

Recent developments in atmospheric chemistry

B. A. Thrush

Department of Physical Chemistry, University of Cambridge, Lensfield Road, Cambridge, UK

The chemistry of the atmosphere was a largely neglected topic until the last decade during which concern about pollution has produced a flurry of activity. The processes which may deplete the relatively small amount of ozone in the stratosphere which protects us from harmful UV light have much in common with those in the Los Angeles-type smog which generate ozone and other oxidants. Such systems involve chain reactions in which free radicals such as HO, HO_2, ClO can carry out a chemical conversion process (ozone to oxygen) many times before they are destroyed. Recent research shows that nitrogen oxides from the exhausts of supersonic aircraft or derived from increased use of nitrogenous fertilisers will have little effect on stratospheric ozone but the potential effects of chlorine compounds, such as aerosol propellants, require careful examination. In contrast, the oxidation of carbon monoxide is now thought to be a major source of ozone in the lower atmosphere.

THE past decade has seen a tremendous growth in research on atmospheric chemistry; this has stemmed largely from the need to identify the processes which remove many of the compounds which man releases into the atmosphere on a large scale and to assess their potential effect on the environment. These problems highlighted our lack of knowledge of the concentrations of important trace species in the unperturbed atmosphere and of the processes which control them. Fortunately both atmospheric and laboratory measurements on these trace species have been greatly aided by recent developments in highly sensitive spectroscopic methods for measuring their concentrations, notably fluorescence, high resolution absorption spectroscopy and laser magnetic resonance.

Computer modelling provides an essential link between laboratory measurements of reaction rates of atmospheric species and the determination of their concentrations in the atmosphere, as the number of reactions is large and the atmospheric circulation a formidably complex problem. Because, in the lower and middle atmosphere, gases are transported by the movement of air masses rather than by molecular diffusion, it will be several years before the massive computational problem of combining a general circulation model of the whole atmosphere with an adequate description of the chemistry can be achieved. In most current work, the transport is parameterised in terms of an 'eddy transport coefficient' which has the same dimensions as a diffusion coefficient.

Two-dimensional models of the atmosphere in which altitude and latitude are variables are now replacing the one-dimensional models involving only altitude; these cannot, for instance, describe the transport of ozone from its predominant formation regime near the equator to the poles where its concentration is greatest.

Comparisons of the measured and calculated concentrations of trace species in the atmosphere provide the best test of the adequacy of our understanding of atmospheric chemistry and transport. Because the tropopause (the temperature minimum at an altitude of about 15 km) provides an effective barrier to transport between the dry, relatively stable stratosphere and the moist rapidly mixed troposphere below it, it is convenient to consider the chemistry of these regions separately (Fig. 1).

0028–0836/78/0276—0345$01.00

© Macmillan Journals Ltd 1978

Stratospheric chemistry

Here the dominant interest is in the processes which control the small amount of ozone which protects the Earth's surface from ultraviolet light with wavelengths less than 300 nm; this corresponds to a 3-mm thick layer of pure ozone gas at ground level. The basic mechanism for the formation and removal of ozone was proposed by Chapman[1] in 1930. Molecular oxygen is dissociated by light with wavelengths less than 242 nm to form oxygen atoms. These combine rapidly with molecular oxygen to form ozone, a process which requires a third body (M) to remove excess energy

$$O_2 + h\nu(\lambda < 242\ \text{nm}) = O + O \quad (1)$$

$$O + O_2 + M = O_3 + M \quad (2)$$

Ozone is very rapidly photolysed by visible and ultraviolet light to regenerate oxygen atoms

$$O_3 + h\nu = O_2 + O \quad (3)$$

It is also removed by reaction with atomic oxygen to regenerate molecular oxygen

$$O + O_3 = O_2 + O_2 \quad (4)$$

There is a fundamental difference between these two decomposition processes for ozone. Process (3) followed by reaction (2) merely interconvert O and O_3 which are termed 'odd oxygen' species, whereas (4) converts the odd oxygen formed by process (1) back to (even) molecular oxygen. Although Chapman himself and others had clearly envisaged that other processes could destroy much atmospheric ozone, only in the present decade have laboratory and atmospheric measurements become accurate enough to establish that reaction (4) only accounts for about one-fifth of the odd oxygen destruction in the atmosphere[2].

The other mechanisms for ozone removal can most easily be envisaged in terms of catalytic cycles in which a species converts O and O_3 into O_2 without being destroyed itself

$$X + O_3 = XO + O_2 \quad (5)$$

$$XO + O = X + O_2 \quad (6)$$

$$O + O_3 = O_2 + O_2 \quad (5)+(6)$$

For reactions (5) and (6) to be rapid, X needs to have an unpaired electron; the importance of atmospheric cycles with X = H, HO, NO and Cl have been recognised[3–6]. (The stability of HF in the atmosphere prevents F contributing to such cycles; there is insufficient bromine in the atmosphere for X = Br to be important.) The hydrogen radicals, H, HO and HO_2, which participate in the X = H and X = HO cycles come mainly from the attack on water vapour of excited oxygen atoms, $O(^1D)$, formed by the photolysis of ozone at wavelengths below 310 nm. These radicals are removed by mutual combination reactions such as

$$HO + HO_2 = H_2O + O_2 \quad (7)$$

$$HO_2 + HO_2 = H_2O_2 + O_2 \quad (8)$$

Figures 2 and 3 illustrate the basic corresponding cycles for X = NO and X = Cl (ref. 7). Here man-made compounds may contribute to ozone destruction. Increased use of nitrogenous fertilisers could lead to significantly increased levels of nitrous oxide in the atmosphere, but the natural sources and sinks of nitrous oxide (N_2O) in the troposphere are not well enough known for an assessment of this effect. The amounts of NO and NO_2 in the stratosphere (totalling about 1 part in 10^8) are also enhanced by their presence in the exhausts of high flying aircraft, but 'rain out' in the troposphere prevents the vastly greater quantities of NO and NO_2 released by combustion processes at or near ground level from reaching the stratosphere. Even

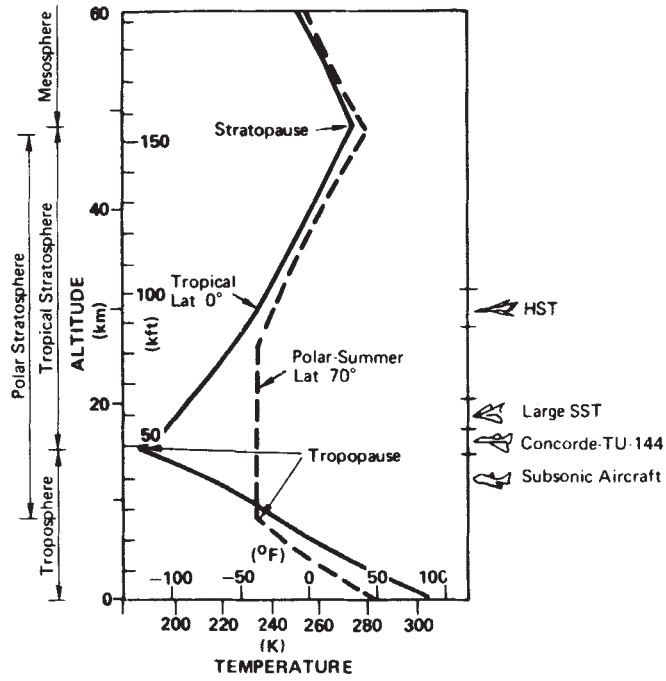


Fig. 1 Temperature of the atmosphere as a function of altitude.

allowing for the fact that the efficient decomposition of NO_2 by near ultraviolet sunlight and the subsequent reaction (2)

$$NO_2 + h\nu(\lambda < 400\ \text{nm}) = NO + O \quad (9)$$

$$O + O_2 + M = O_3 + M \quad (2)$$

$$NO_2 + O_2 + h\nu = NO + O_3 \quad (9)+(2)$$

effectively reverse the ozone destruction process

$$NO + O_3 = NO_2 + O_2 \quad (10)$$

the nitrogen oxides provide the strongest removal process for odd oxygen. However, man's efforts in adding to the stratospheric burden of nitrogen oxides or chlorine compounds cannot be assessed directly from this, as recent laboratory studies[8,9] have established the rapidity of the reaction

$$HO_2 + NO = HO + NO_2 \quad (11)$$

which, like the closely related process

$$ClO + NO = Cl + NO_2 \quad (12)$$

does not itself remove odd oxygen, but links the odd oxygen destruction cycles so that the effects of nitrogen oxides, halogens, and so on, on ozone are not additive. On this basis it is believed that current and forecast amounts of stratospheric flight (for example, by Concorde and Boeing SP747) will have a negligible effect on stratospheric ozone, but the effects of chlorine compounds are magnified[9–11].

The present levels of chlorine in the stratosphere (1.5–2 parts in 10^9) arise mainly from methyl chloride[12] (from algae and from slash-and-burn agriculture), from salt spray and from volcanic gases. These sources will soon be overtaken by the chlorofluoromethanes $CFCl_3$ and CF_2Cl_2 now present at 1 and 2 parts in 10^{10} respectively in the troposphere[13]. These compounds, which are used as aerosol propellants, refrigerants and foam-blowing agents, have no known natural removal process other than transport into the stratosphere followed by decomposition to yield chlorine atoms, and eventual removal as hydrogen chloride, a process which takes 5–10 years. However, 90% of the halocarbon will remain in the troposphere under uniform mixing conditions because this region contains 90% of the molecules in the atmosphere. The atmospheric lifetimes of $CFCl_3$ and CF_2Cl_2 are therefore believed to be 50–100 years, over which period the depletion of ozone is predicted to increase

to near 15% if these substances continue to be released at current rates. However, uncertainties in the rate coefficients of the 50–100 reactions involved and in the transport parameters each contribute a factor of two to the uncertainty of such estimates[7,10]. Another source of uncertainty is the 'unknown chemistry' involving perhaps compounds such as HO_2NO_2, $ClONO_2$ and HOCl which have very limited stability in the laboratory but might be present in the stratosphere in amounts approaching the current detection limits (about 1 part in 10^9).

Verification of the model calculations depends on reliable measurements of trace species in the atmosphere which, with the exception of ozone, are few in number and show significant fluctuations apparently associated with the past history of the air masses sampled. For this reason, simultaneous measurements of a number of trace species are now considered particularly important, as they can provide a better test of the correct identification of the reactions interconverting these species. For instance, the ratio of NO_2 to NO at 20–25 km altitude is predominantly controlled by the rates of reactions (9), (10) and (13)

$$O + NO_2 = NO + O_2 \qquad (13)$$

giving

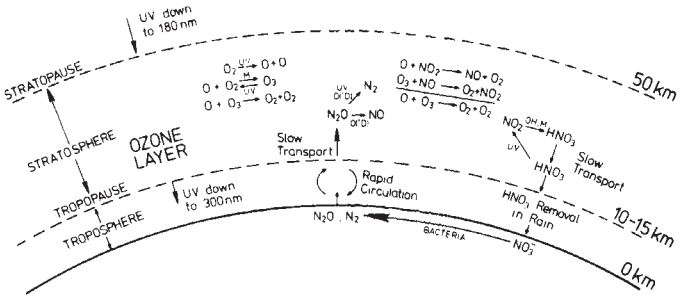
$$\frac{[NO_2]}{[NO]} = \frac{k_{10}[O_3]}{J_9 + k_{13}[O]}$$

the ratio of $[O_3]$ to [O] being governed by processes (2) and (3). Thus, the measurement of the ratio $[NO_2]/[NO]$ at these and at higher altitudes where reactions (11) and (12) become significant can verify the accuracy of our understanding of stratospheric chemistry. For these species, the measurements are normally carried out by long path infrared spectroscopy of the setting sun when the ratio of NO_2 to NO is increasing rapidly because the photolysis of O_3 and NO_2 is declining markedly and agreement with the model is not yet perfect[14]. Anderson *et al.*[15] has measured Cl and ClO concentrations at altitudes of 25–40 km by balloon flying and parachuting a flow apparatus in which chlorine atoms are detected in fluorescence either as initially present or produced by reaction (12). The values which he finds are too high to be readily reconciled with measurements of HCl at lower altitudes (with which it should be in steady state, as indicated in Fig. 3) or with known ozone concentrations. This and evidence for a seasonal variation make simultaneous measurements of Cl, ClO, HO and O or O_3 a high priority experiment.

Tropospheric chemistry

Here, as with stratospheric chemistry, much of the recent research has been governed by pollution problems. Historically, sulphur dioxide has been the most important air pollutant in the UK and although its concentration in urban air has been greatly reduced by the use of alternative fuels and by dispersal through high chimneys, the mechanism of its oxidation over a period of weeks before its deposition as sulphuric acid in rain or as sulphate aerosol is far from understood. Homogeneous reactions involving combination reactions with HO in unpolluted

Fig. 2 Dominant ozone formation and removal processes in the natural atmosphere.



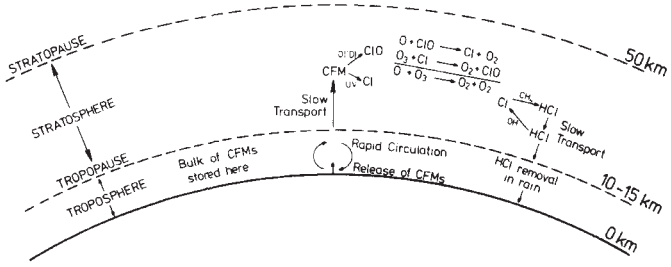


Fig. 3 Simplified diagram of the behaviour of chlorofluoromethanes in the atmosphere.

and HO_2 in polluted atmospheres are believed to be important and there is evidence for a catalytic process involving transition metals, notably manganese, in polluted atmospheres[16].

With an increase in the use of cars, the place of the traditional London fog is beginning to be taken by photochemical smogs which have been a familiar problem in the USA for at least 20 years. These occur when a temperature inversion in sunny weather prevents the dispersal of the exhaust gases of internal combustion engines. The photolysis of NO_2 (reaction (9) above) produces oxygen atoms, these and the ozone which they produce by combination with molecular oxygen (reaction (2)) attack particularly the unsaturated hydrocarbons which are also present in car exhausts. This initiates a highly complex series of oxidation reactions involving particularly the free radicals HO and HO_2 which yield *inter alia* aldehydes and the hitherto unknown peroxy-acyl nitrates ($RCO.OONO_2$), which like the ozone also present are powerful irritants. Carbon monoxide is also oxidised to carbon dioxide. Computer models of the very complicated series of reactions which are believed to occur in smog can now reproduce the smog in laboratory chambers from known materials and also the diurnal variation of the main species involved in the atmosphere, but many individual steps require laboratory measurements[17]. A recent development has been the detection of nitrosamines in Los Angeles smog when sources of aliphatic amines are present[18].

Anthropogenic sources of ozone may well be important even in apparently unpolluted regions of the lower atmosphere. Fishman and Crutzen[19] have recently pointed out that the preponderance of land in the Northern Hemisphere results in three times as much ozone destruction at the surface there compared with the Southern Hemisphere but there is no evidence of a corresponding ratio of ozone fluxes downwards from the stratosphere; this would be required if this were the main ozone source. They conclude that there must be significant tropospheric sources of ozone in the Northern Hemisphere and suggest that these are anthropogenic: the oxidation of hydrocarbons and the oxidation of carbon monoxide

$$CO + HO = CO_2 + H \qquad (14)$$

$$H + O_2 + M = HO_2 + M \qquad (15)$$

$$HO_2 + NO = HO + NO_2 \qquad (11)$$

$$NO_2 + h\nu = NO + O \qquad (9)$$

$$O + O_2 + M = O_3 + M \qquad (2)$$

giving

$$CO + 2O_2 = CO_2 + O_3$$

As in the stratosphere, the HO radicals which catalyse these oxidations are produced by the reaction with water vapour of a small portion of the excited oxygen formed in the photolysis of ozone at short wavelengths

$$O_3 + h\nu(\lambda < 310\ \text{nm}) = O_2 + O(^1D) \qquad (16)$$

$$O(^1D) + H_2O = HO + HO \qquad (17)$$

Hydroxyl is the most reactive trace species in the troposphere and is therefore the dominant scavenger of many anthropogenic substances such as CO, hydrocarbons and halocarbons containing C—H and C=C linkages. Its concentration in the unpolluted troposphere is so low that it has proved very difficult to measure,

but it is currently believed to be about 3 parts in 10^{14} at ground level, which may be compared with a value of 3 in 10^{10} in the stratosphere above 40 km where it contributes significantly to ozone removal.

Reaction with HO gives estimated tropospheric lifetimes of 2 weeks and 10 yr respectively for the important solvent molecules $CHCl{=}CCl_2$ and $CH_3.CCl_3$, which is progressively replacing it. In fact measurement of the atmospheric concentration of a compound such as $CH_3.CCl_3$ with well-known and wholly anthropogenic sources may well provide the best method for determining the average concentration of HO in the troposphere and thus its ability to remove other species from the lower atmosphere. CCl_4 is, like $CFCl_3$ and CF_2Cl_2, present in the troposphere at the 1–2 parts in 10^{10} level[13], being also unreactive towards HO. There is no obvious natural source or tropospheric removal process for CCl_4 and, as this substance is now mainly used as a chemical intermediate, it is plausible that the present quantities in the atmosphere have persisted from the time when it was used extensively as a solvent.

In an article of this length it is only possible to touch on a limited number of the topics of current interest in atmospheric chemistry. Such important but complex problems as the evolution of the Earth's atmosphere and natural cycles of such species as nitrous oxide have been omitted in favour of areas on which there have been particular emphasis in the past few years.

1. Chapman, S. *Mem. R. M. Soc.* **3,** 103–125 (1930).
2. Johnston, H. S. *A. Rev. phys. Chem.* **26,** 315–338 (1975).
3. Bates, D. R. & Nicolet, M. *J. geophys. Res.* **55,** 301–327 (1950).
4. Crutzen, P. J. *Q. Jl R. met. Soc.* **96,** 320–325 (1970).
5. Johnston, H. S. *Science* **173,** 517–522 (1971).
6. Molina, M. J. & Rowland, F. S. *Nature,* **249,** 810–812 (1974).
7. *Halocarbons: Their Effect on Stratospheric Ozone* (US natn. Acad. Sci., Washington, 1976).
8. Howard, C. J. & Evenson, K. M. *Geophys. Res. Lett.* **4,** 437–440 (1977).
9. Thrush, B. A. *Phil. Trans. R. Soc. A.* (in the press).
10. *Chlorofluoromethanes and the Stratosphere* (NASA RP-1010, 1977).
11. Fabian, P. *Nature* **272,** 306–307 (1978).
12. Lovelock, J. E. *Nature* **256,** 193–194 (1975).
13. Singh, H. B., Salas, S., Shigeishi, H. & Crawford, A. *Atmos. Env.* **11,** 819–828 (1977).
14. Evans, W. F. J. *et al. Atmosphere* **14,** 189–198 (1976).
15. Anderson, J. G., Margitan, J. J. & Stedman, D. H. *Science* **198,** 501–503 (1977).
16. *Pollution in the Atmosphere: A Study Group Report* (Royal Society, London, 1978).
17. Kerr, J. B., Calvert, J. G. & Demerjian, K. L. *Chem. Brit.* **8,** 252–257 (1972).
18. Pitts, J. N. *Phil. Trans. R. Soc. A* (in the press).
19. Fishman, J. & Crutzen, P. J. *Nature* **274,** 855–858 (1978).

Solar–terrestrial influences on weather and climate

George L. Siscoe

Department of Atmospheric Sciences, University of California, Los Angeles, California 90024

During the past century over 1,000 articles have been published claiming or refuting a correlation between some aspect of solar activity and some feature of terrestrial weather or climate. Nevertheless, the sense of progress that should attend such an outpouring of 'results' has been absent for most of this period. The problem all along has been to separate a suspected Sun–weather signal from the characteristically noisy background of both systems. The present decade may be witnessing the first evidence of progress in this field. Three independent investigations have revealed what seem to be well resolved Sun–weather signals, although it is still too early to have unreserved confidence in all cases. The three correlations are between terrestrial climate and Maunder Minimum-type solar activity variations, a regional drought cycle and the 22-yr solar magnetic cycle, and winter hemisphere atmospheric circulation and passages by the Earth of solar sector boundaries in the solar wind. The apparent emergence of clear Sun–weather signals stimulated numerous searches for underlying physical causal links.

ACTIVE interest in the field of Sun–weather relationships is probably higher at the present time than it has ever been in the past. Since 1972 there have been six major national or international conferences on the subject. The literature is correspondingly very large, and readers interested in going beyond what will be emphasised in this short review should consult the proceedings that have (or will have) emerged from some of these conferences[1–3], and also the extensive references contained in a recent review by Pittock[4].

Solar variability

The subject of Sun–weather relationships is built out of correlations, real or apparent, between changes in solar activity and changes in terrestrial weather and climate. No matter how variable might be the terrestrial atmosphere, if the Sun did not vary, there would be no such subject. It is therefore useful to begin with a brief account of solar variations. These fall naturally into distinct time divisions, the most familiar of which is the 11-yr sunspot cycle. There is also a 22-yr oscillation in the magnetic polarity of the leading member of sunspot pairs, the so-called Hale magnetic cycle. A longer variation of very roughly 80 yr, referred to as the Wolf–Gleissberg cycle, is seen in sunspot cycle amplitudes, as measured for example by the annual mean sunspot number.

On an even longer timescale, there is an important, recently rediscovered kind of solar variation which is typified by the now-famous Maunder Minimum, an interval in the history of solar activity extending approximately from AD 1645 to 1715 (ref. 5). The Maunder Minimum is remarkable for the near-complete absence of spots on the face of the Sun, and a concurrent absence from terrestrial skies of the aurora, a good local indicator of solar activity. It is now generally believed that ^{14}C anomalies detected in tree rings of known, specific ages mark times in the past when other weak solar activity episodes have occurred[6,7]. Episodes of anomalously strong solar activity are indicated as well. The intervals between anomalies vary, but are perhaps characteristically around 400 yr. There is also a hint of a slower variation with a period around 2,500 yr associated with clustering of anomalies of the same sign.

On timescales less than the 11-yr period of the sunspot cycle, there are annual and semi-annual variations in geomagnetic and auroral activity resulting from orbital variations in the efficiency with which solar activity energises the Earth's magnetosphere. On still shorter timescales, there is the 27-d period of the solar

© Macmillan Journals Ltd 1978

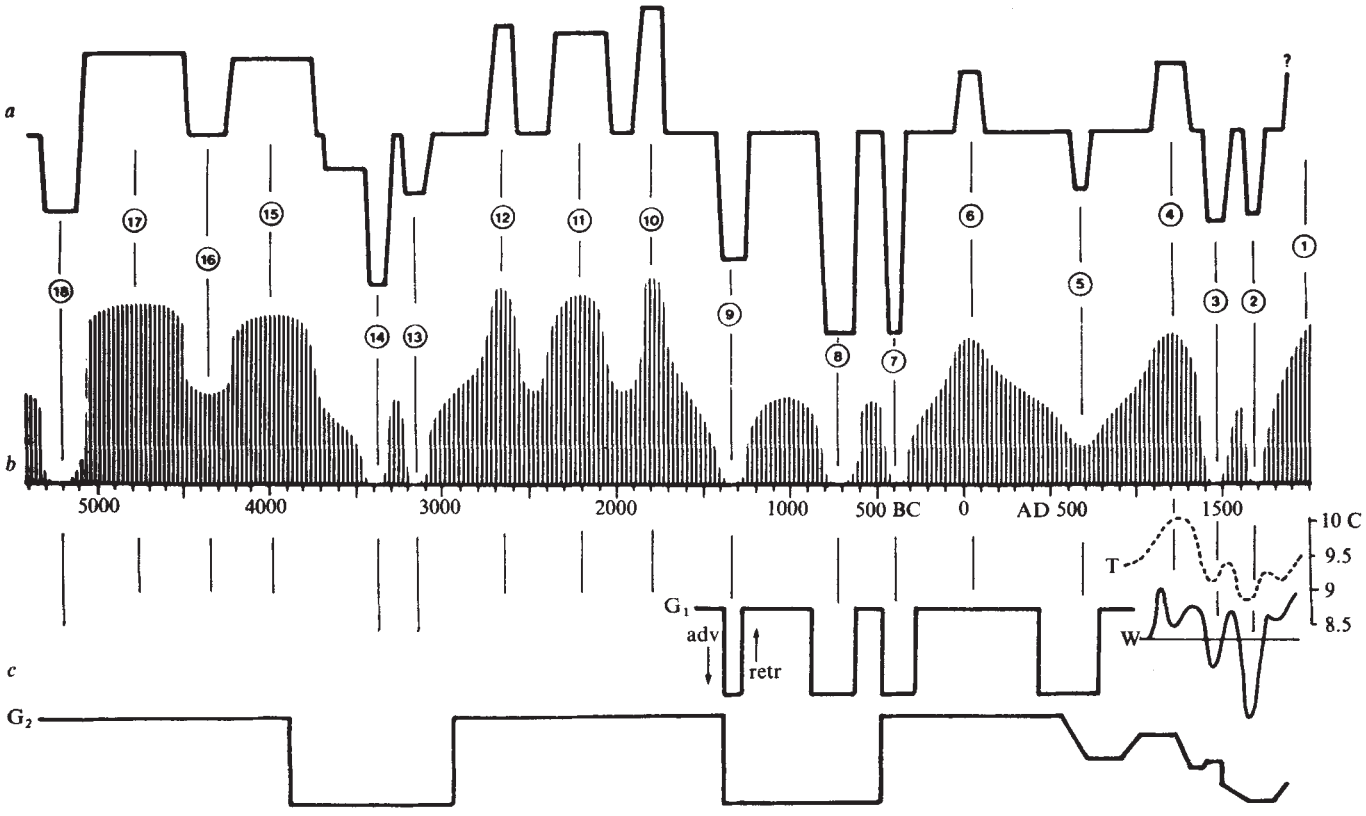


Fig. 1 From Eddy[7]. *a*, Persistent deviations in ^{14}C plotted schematically and normalised to feature 2 (Maunder Minimum); downward excursions refer to increased relative ^{14}C and imply decreased solar activity. *b*, Interpretation of curve *a* as a long-term envelope (of possible sunspot cycle) that minimises in features 2, 3, and so on, and maximises at 4, 6, and so on. *c*, Four estimates of past climate. Step-curve G_1: times of advance and retreat of Alpine glaciers, after Le Roy Ladurie[16]. Curve G_2: same for worldwide glacier fluctuations, from Denton and Karlen[17]. Curve T: estimate of mean annual temperature in England (scale at right), after Lamb[18]. Curve W: winter-severity index for Paris–London area, from Lamb[19], downward is colder.

rotation, and the variable, but approximately 8-d spacings between passages by the Earth of solar sector boundaries. In addition, at irregular intervals, solar flares produce geomagnetic storms which persist typically for 1–2 d.

Terrestrial atmospheric responses to all of the solar activity variation types listed above have been claimed or suggested, with the exception of the annual and semi-annual variations which would be impossible to detect against the background of the yearly succession of the seasons. A panel composed of participants at the symposium held recently at Ohio State University came to the historically unprecedented conclusion that three of the claimed correlations must now be granted the status of reasonable to virtual certainty. These three, correlations with Maunder Minimum-type episodes, the Hale magnetic cycle, and sector boundary passages, will be emphasised in the limited space available here. Other possible correlations, such as with the sunspot cycle and with the solar rotation, however, retain the controversial status that has traditionally characterised claims of Sun–weather correlations.

Maunder Minimum-type episodes

Atmospheric responses to the kind of changes in the Sun for which the Maunder Minimum is archetypal are of the clearest sort, global changes in climate[5–7]. The Maunder Minimum itself together with the preceding Spörer Minimum (AD 1420–1570) coincide with the two coldest excursions within the Little Ice Age, a period encompassing approximately the fifteenth to seventeenth centuries, when the surface temperatures in Europe and America were depressed by 0.5–1°C. In the case of the Maunder Minimum, the climatic event is linked directly to anaomalous solar behaviour through an ample record of telescopic observations of the Sun covering the period in question. Earlier climatic excursions, such as the Spörer Minimum and the Mediaeval Climatic Optimum, a globally warm period during the twelfth century, are linked to solar changes through proxy data. Foremost among these in importance is the tree-ring record of ^{14}C anomalies, which are believed to result when a change in the level of solar activity varies the flux of ^{14}C-producing galactic cosmic rays that can penetrate the Solar System to Earth's orbit. High solar activity reduces the galactic cosmic ray flux at Earth, which depresses the rate of ^{14}C production. The opposite condition characterises periods of attenuated solar activity. When the curve of tree ring ^{14}C anomalies is compared with the curve of climatic variations, in the words of J. A. Eddy, who has convincingly promoted and applied the field of long-term solar changes, "the correspondence, feature for feature, is almost the fit of a key in a lock" (see Fig. 1)[7].

Before the ^{14}C studies, the same long-term Sun–climate correlation had been recognised by J. R. Bray[8,9], using the solar activity index constructed by J. D. Schove[10] from historical observations of aurorae and naked-eye sunspots. A more recent and more extensive compilation of historical aurorae which combines occidental and oriental observations also shows the existence of major solar activity features corresponding to the major climatic excursions[11]. Although the presently known auroral record becomes too fragmentary before the first few centuries of the Christian Era to identify solar activity episodes of the Maunder type, it may serve to validate for that purpose the use of the much longer record of tree-ring ^{14}C anomalies.

Hale magnetic cycle

The correlation that has been found linking atmospheric behaviour and the (approximately) 22-yr solar magnetic cycle involves a modulation of regional climate. The Great American Desert, which is centred roughly in the southwestern US and northern Mexico, has long been known to expand and contract its area of influence with a timescale of about 20 yr producing cycles of drought in the region west of the Mississippi River. The

'dust bowl' of the 1930s and the droughts of the 1950s and 1970s are the most recent of its intensifications and expansions. Many people have suspected and presented evidence to show that the drought cycle is related to the Hale double sunspot cycle[12-14]. Until recently the record length was too short and the drought cycle was not quantified, so that statistical tests of the proposed correlations were difficult to perform.

Mitchell and Stockton have now succeeded in quantifying the temporal climatic behaviour of the desert with a single parameter, yearly index extending back to at least AD 1700 constructed out of tree-ring indices from 40 regionally distributed, climatically sensitive sites[15]. (Actually three families with four sets of such indices were generated to establish internal consistency of the procedure.)

Power spectral analyses of the obtained data sets, so-called drought area indices (DAI), revealed well defined concentrations of power at periods near 22 yr, at formal significance levels ranging from 95% to more than 99%. The analyses were based on sets of 262 data points from the time interval 1700–1962, which included approximately 12 complete Hale cycles. Phase locking between the succession of Hale magnetic cycles and the cycles of drought area was demonstrated to exist at formal significance levels exceeding 99%. Thirdly, the amplitudes of the drought cycles vary from one to the next in a way that bears a strong resemblance to the envelope of the amplitudes of the Hale sunspot cycle. The linear correlation between the two data sets is statistically significant at between the 95 and 99% levels. Weak sunspot cycle amplitudes tend to correspond to minimum drought conditions, but the authors are quick to point out that the apparent correlation is not a reliable basis for operational climate prediction.

Although Pittock[4] has urged caution in accepting at face value the inferred Sun–weather correlation because "complex questions are involved in assessing the true statistical significance of the result", the Hale solar cycle–regional climate connection must be considered at least a serious possibility in view of its having survived the three levels of statistical tests. Full confidence in the result will most likely be achieved only if it is independently confirmed by other researchers who, like Mitchell and Stockton, have a critical attitude toward testing its reality. Such confirmation has been secured by the third of the cited correlations, which is described next.

Solar sector boundary passages

A line of investigation focused on the possibility that geomagnetic activity (a solar effect) may influence atmospheric circulation emerged in the 1950s (refs 20–23). It evolved in cast and sophistication under the stimulus of consistently positive findings until a rigorous statistical test of a specific form of the hypothesis recently could be made[24,25]. The crucial point was reached when two operationally specified, objective data sets were identified, one representing atmospheric circulation, and the other containing a statistically significant number of 'solar' events that could be used for the purposes of a correlation analysis.

The solar input to the correlation is a list of the times of passages by the Earth of solar wind sector boundaries. Sectors are regions in the solar wind, first recognised by Wilcox and Ness[26], that have predominantly one magnetic polarity, either toward or away from the Sun, and they are separated from adjacent sectors of opposite polarity by thin, usually well defined boundaries. Magnetic sectors and their boundaries tend to co-rotate with the Sun. Normally there are either two or four sectors at any given time. Magnetic sector boundaries have been shown to be natural time markers, that bear a temporally fixed relationship to the progression of a variety of solar wind phenomena. By mapping them back to their origins on the Sun, they become also natural space markers for solar features.

The atmospheric input to the correlation is a derived quantity called the vorticity area index (VAI) first defined by Roberts and Olson[24]. On a map of a constant pressure surface in the atmosphere (usually the 300 mbar level, but the correlation is also evident at 500 mbar), absolute vorticity contours are drawn at selected vorticity levels. The VAI is based on the portion of the map for the Northern Hemisphere north of 20° N, and is defined as the sum of the area (in square kilometres) over which the absolute vorticity exceeds $20 \times 10^{-5}\ s^{-1}$ plus the area over which it exceeds $24 \times 10^{-5}\ s^{-1}$. The index is specifically designed to measure the intensity and extent of low pressure troughs, which may roughly be thought of as large 'bad weather' systems.

Wilcox *et al.* have carried out a superposed epoch analysis on the VAI, using as zero days the times in a previously published list of solar sector boundary passages[27-32]. The results of their analysis based on 54 boundaries, reproduced in Fig. 2, shows an evidently statistically significant feature, of about a 6-d temporal width, centred approximately 1 d after the boundary passage, at which time the vorticity area index is depressed about 10% relative to adjacent values. The apparent signal was clearly evident only for the winter months and was entirely absent for the summer months.

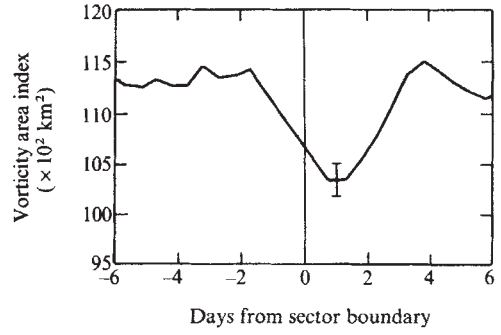


Fig. 2 From Wilcox *et al*[27]. Average response of the vorticity area index (the area of all the low pressure troughs in the Northern Hemisphere) based on the times of 54 solar magnetic sector boundary passages by Earth.

The relatively short timescale of this particular claimed Sun–weather effect, less than 2 weeks, and the comparatively long interval over which VAI data were available, many years, permitted extensive testing of the statistical properties of the index over the appropriate timescale. Hines and Halevy[33,34] have undertaken the task of searching for an explanation of the alleged correlation in terms of accidental properties of the two data sets (such as having a common period of variation but derived from known, independent causes) but none was found. They ultimately concluded that either the correlation was real or it was a random fluctuation for which their analysis of the data provided a formal *a priori* probability of one chance in 20. Shapiro has carried out a similar analysis and reached the same conclusion[35]. Both critics of the claimed correlation pointed out, however, that the *a priori* probability is really not applicable because the data were in some respects selected *a posteriori*, only winter months, only a certain height interval in the atmosphere, and so on.

The issue was decided in favour of the reality of the correlation when a list of times of additional boundary passages was compiled, containing more than the original number of entries. Only if the correlation were real would one expect the signal seen in Fig. 2 to reappear in essentially identical form from a superposed epoch analysis based on the new list of times. After performing the critical test and discovering the same signal, Hines and Halevy were forced to conclude that "the alleged solar signal should be accepted as a physically meaningful signal whose origins should now be sought"[34,36].

Other correlations

Within the past century over 1,000 articles have been published asserting or refuting a claim of some form of Sun–weather influence. Most of these are concerned with correlations with the 11-yr sunspot cycle. King has published a veritable catalogue of such claimed correlations[37]. The whole field of 11- and 22-yr

correlations has been critically discussed in a recent, thorough review by Pittock[4]. Multiple claims of sunspot correlations have been made for atmospheric pressure, temperature, precipitation, thunderstorms, tropopause height, and ozone, and isolated claims for an assortment of other weather–climate variations. Pittock, while reserving judgment on the new results on the drought cycle in the Western United States, concludes for the rest that "despite a massive literature on the subject, there is at present little or no convincing evidence of statistical significant or practically useful correlations between sunspot cycles and the weather or climate". Some advocates in the Sun–weather community are likely to disagree.

When statistical tools are needed to determine the reality of a correlation, especially a non-predictable one, absolute certainty will never be achieved. Instead, levels of conviction are achieved, and for the same statistical data, these vary between individuals according to whether they are advocates, agnostics or critics. Pittock's conclusion tells us that there is not yet in the field of sunspot–weather relations a single correlation with enough statistical punch to convince a critic of its reality.

It might turn out that some of the correlations are real, but a common problem that weakens the statistical approach is the shortness of the atmospheric data record compared to a signal period of 11 or 22 yr. In some cases it may be possible to extend the data record, for example through the use of proxy data, as was done for the drought cycle, with a resultant considerable increase in the strength of the statistical evidence.

As an alternative solution to the short data record problem, correlation seekers have been advised by Roberts[38] to concentrate more on short-term responses of the atmosphere to solar activity variations. The sector boundary correlation with the vorticity area index stands as a model of what can be achieved in this way. Geomagnetic and solar surface variables have been found to correlate with the VAI as well[39]. Other short-term correlations have already been suggested or claimed, such as between sunspots and quasi-biennial variations in the stratosphere[40], between the solar rotation and large amplitude standing planetary waves in the atmosphere[41], and between geomagnetic activity and a behaviour characteristic of persistence in the sea-level pressure distribution[42]. These have yet to be subjected to critical independent review.

One recently claimed short-term correlation needs to be mentioned. Attempts based on correlations with the 11-yr sunspot cycle to determine if there is a variation of global ozone with solar activity have been inconclusive. Keating[43], following a suggestion of London and Oltmans[44], has correlated monthly averaged variations of global ozone, determined from a remote sensing detector aboard the Nimbus 4 satellite, with solar activity as measured by the monthly mean 10.7 cm solar electromagnetic flux. A correlation exceeding the 95% level of confidence was found, although only 10 months of data were available for the study. The relative amplitude of the ozone variation over the study period was slightly greater than 0.1% compared with the average.

The possibility that some property of atmospheric ozone may be influenced by solar activity, which now needs to be independently confirmed or refuted with different data sets, is especially important because of the central role that ozone modulations play in current thinking about mechanisms for Sun–weather connections. It seems clear that the subject of Sun–weather connections will have the appearance of an underdeveloped science until at least one physical causal link (or one complete chain of causal links) has been verified. The determination of operating physical mechanisms would then provide a powerful means, for example, through inclusion in numerical climate models, to answer questions about the solar activity influences on specific weather–climate behaviour.

Mechanisms

It is generally recognised that there may be more than one cause and more than one effect involved in the total atmospheric response to the complex of variations that make up solar activity. The most obvious possibility, solar activity-related variation in the solar constant which powers the weather machine, remains an attractive possibility. This is true especially for the longer-term correlations, even though a recent assessment concludes that the solar constant has remained unchanged to within 0.75% over the last half of the previous solar cycle[45]. An analysis by Volland finds that planetary-scale pressure waves driven by a 0.1% 27-d oscillation of the solar constant would have detectable amplitudes[46,47]. In view of the simple direct consequences of small variations in the solar constant, there is a consensus that long-term, high-precision measurements of it would be very desirable.

Regarding Maunder Minimum-type variations, Eddy points out again that the simplest and most straightforward explanation is that solar behaviour includes a series of long-term changes in the solar constant, in this case with an amplitude of the order of perhaps 1% (refs 6 and 7). In his view, solar constant variations modulate the envelope of sunspot cycle amplitudes, accounting for the observed correlations with solar activity.

Other possibilities for explaining long-term correlations nevertheless exist. Changes in the thermal structure of the stratosphere are shown by Bates to be capable of significantly affecting the ability of the lower atmosphere to transport heat from the Equator to the Pole[48]. The thermal structure of the stratosphere can be influenced by solar activity through its modulation of solar ultraviolet radiation, which is stopped in the stratosphere[49,50], and through its supposed modulation of stratospheric ozone, which absorbs part of the solar ultraviolet. Another avenue of influence connects the same solar activity modulation of galactic cosmic rays that produces the ^{14}C anomalies in tree rings with the tendency of such cosmic rays to deplete ozone in the stratosphere, indirectly through a chain of stratospheric chemistry[51,52].

In spite of these promising leads concerning long-term correlations, it seems fairly obvious that explanations in terms of variations of one or another solar electromagnetic flux will not account for the VAI correlation with sector boundary passages. If we adopt the position expressed by Hines and Halevy, that it is time to search for the mechanisms behind this correlation, we are faced with the prospect that the responsible agents are probably part of the complex of solar particle fluxes, including the solar wind and solar cosmic rays; in which case the complete story of causal linkages will have to include the interaction of these particles with Earth's magnetosphere, and the subsequent interaction of the magnetosphere with the atmosphere.

Investigators attempting to discover such linkages have been adequately warned that the solar particle fluxes contain negligible energy compared to weather systems with which they are supposedly correlated[53,54]. On the other hand two studies show that the sector boundary–VAI correlation does not involve a detectable change in atmospheric energetics[55,56]. The situation thus favours the kinematical explanation offered by Hines and Halevy, in which phase shifting of pre-existing meteorological noise is invoked to produce the observed coherence[33,34]. The required phase-shift mechanism, however, has not yet been identified.

Present exploration aimed at locating a coupling mechanism that can convert a solar wind stimulus into an atmospheric response on a short timescale is focusing on the magnetospheric particles that precipitate into the upper atmosphere, and which are known to be modulated by the solar wind. Examples are the precipitations that produce the polar aurorae[57], and the rarer but more penetrating "relativistic electron precipitation" (REP) events[58]. Some argue that these may act, for example, through promoting the development of cirrus clouds high in the troposphere[24] or by modifying the elements in the global atmospheric electrical circuit, thereby possibly affecting thunderstorm activity[59].

This short list of ideas is not intended to show that the problem of identifying the causal links connecting solar activity and weather or climate is close to being solved. On the other hand, now that some of the claimed correlations have risen respectably

high out of the noise, the problem is receiving increasingly more attention. There is reason for optimism that the next landmark in the field of Sun–weather relationships, the verification of a causal link, may not be far off.

1. Bandeen, W. R. & Maran, S. P. (eds) *Possible Relationships between Solar Activity and Meteorological Phenomena* (NASA SP-366, Washington, DC, 1975).
2. Johnson, F. S. (ed.) *Collection of Extended Summaries of Contributions Presented at Joint IAGA/IAMAP Assembly*, Seattle, Washington, Joint Symposium C, 22 Aug.–3 Sept., 1977 (NCAR, Boulder, Colorado, 1978).
3. *Proceedings of Ohio State University Symposium/ Workshop on Solar Terrestrial Influences on Weather and Climate*, July 24–28, 1978 (in the press).
4. Pittock, A. B. *Rev. Geophys. Space Phys.* **16,** 400 (1978).
5. Eddy, J. A. *Science* **192,** 1189 (1976).
6. Eddy, J. A. in *Physics of Solar Planetary Environments* 958 (Am. Geophys. Union, Washington, DC, 1976).
7. Eddy, J. A. *Climatic Change* **1,** 173 (1977).
8. Bray, J. R. *Science* **168,** 571 (1970).
9. Bray, J. R. *Science* **171,** 1242 (1971).
10. Schove, D. J. *J. geophys. Res.* **60,** 127 (1955).
11. Keimatsu, M. *Ann. Sci. Kanazawa Univ.* **13,** 1 (1976).
12. Borchert, J. R. *Ann. Ass. Am. Geogr.* **61,** 1 (1971).
13. Marshall, J. R. thesis, Univ. Kansas (1972).
14. Thompson, L. M. *J. Soil Water Conserv.* **28,** 87 (1973).
15. Mitchell, J. M., Jr, Stockton, C. W. & Meko, D. M. in *Proceedings of Ohio State University Symposium/ Workshop on Solar-Terrestrial Influences on Weather and Climate*, July 24–28, 1978 (in the press).
16. Le Roy Ladurie, E. *Histoire du Climat depuis l'an mil* (Flammarion, Paris, translated by B. Bray, Doubleday, New York, 1971).
17. Denton, G. H. & Karlen, W. *Quatern. Res.* **3,** 155 (1973).
18. Lamb, H. H. *Climate: Present, Past and Future* **1** (Methuen, London, 1972).
19. Gates, W. L. & Mintz, Y. *Understanding Climate Change*, Appendix A (Natn. Acad. Sci., U.S.A., 1975).
20. Shapiro, R. *J. Meteor.* **13,** 335 (1956).
21. Woodbridge, D., Macdonald, N. J. & Pohrte, T. W. *J. geophys. Res.* **64,** 331 (1959).
22. Macdonald, N. J. & Roberts, W. O. *J. geophys. Res.* **65,** 529 (1960).
23. Macdonald, N. J. & Roberts, W. O. *J. Meteor.* **18,** 116 (1961).
24. Roberts, W. O. & Olson, R. H. *J. atmos. Sci.* **30,** 135 (1973).
25. Roberts, W. O. & Olson, R. H. *Rev. Geophys. Space Phys.* **11,** 731 (1973).
26. Wilcox, J. M. & Ness, N. F. *J. geophys. Res.* **70,** 5793 (1965).
27. Wilcox, J. M., Scherrer, P. H., Svalgaard, L., Roberts, W. O. & Olson, R. H. *Science* **180,** 185 (1973).
28. Wilcox, J. M. *et al. J. atmos. Sci.* **31,** 581 (1974).
29. Wilcox, J. M., Svalgaard, L. & Scherrer, P. H. *Nature* **255,** 539 (1975).
30. Wilcox, J. M. *J. atmos. terres. Phys.* **37,** 237 (1975).
31. Wilcox, J. M. *Science* **192,** 745 (1976).
32. Wilcox, J. M., Svalgaard, L. & Scherrer, Ph. H. *J. atmos. Sci.* **33,** 1113 (1976).
33. Hines, C. O. & Halevy, I. *Nature* **258,** 313 (1975).
34. Hines, C. O. & Halevy, I. *J. atmos. Sci.* **34,** 382 (1977).
35. Shapiro, R. *J. atmos. Sci.* **33,** 865 (1976).
36. Wilcox, J. M. *Science* **192,** 745 (1976).
37. King, J. W. *Astronaut. Aeronaut.* **13,** April, 10 (1975).
38. Roberts, W. O. in *Possible Relationships between Solar Activity and Meteorological Phenomena* 247 (NASA SP-366, Washington, DC, 1975).
39. Olson, R. H., Roberts, W. O., Prince, H. D. & Hedeman, E. R. *Nature* **274,** 140 (1978).
40. Sugiura, M. *Geophys. Res. Lett.* **3,** 643 (1976).
41. King, J. W., Slater, A. J., Stevens, A. D., Smith, P. A. & Willis, D. M. *J. atmos. terr. Phys.* **39,** 1357 (1977).
42. Shapiro, R. *J. atmos. Sci.* **29,** 1213 (1972).
43. Keating, G. M. *Nature* **274,** 873 (1978).
44. Angell, J. K. & Korshover, J. *Mon. Wea. Rev.* **104,** 63 (1976).
45. Willson, R. C. & Hickey, J. R. in *The Solar Output and Its Variation* 111 (Colo. Assoc. Univ. Press, Boulder, Colorado, 1977).
46. Volland, H. *Nature* **269,** 400 (1977).
47. Volland, H. in *Collection of Extended Summaries of Contributions Presented at Joint IAGA/IAMAP Assembly*, Seattle, Washington, Joint Symposium C, 22 Aug.–3 Sept., 1977 (NCAR, Boulder, Colorado, 1978).
48. Bates, J. R. *Q. Jl R. met. Soc.* **103,** 397 (1977).
49. Reck, R. A. *Science* **192,** 557 (1976).
50. Reid, G. C., Isaksen, I. S. A., Holzer, T. E. & Crutzen, P. J. *Nature* **259,** 177 (1976).
51. Ruderman, M. A., Foley, H. M. & Chamberlain, J. W. *Science* **192,** 555 (1976).
52. Chamberlain, J. W. *J. atmos. Sci.* **34,** 737 (1977).
53. Dessler, A. J. in *Possible Relationships between Solar Activity and Meteorological Phenomena* 187 (NASA SP-366, Washington, DC, 1975).
54. Willis, D. M. *J. atmos. terr. Phys.* **38,** 685 (1976).
55. Williams, R. G. *Geophys. Res. Lett.* **5,** 519 (1978).
56. Bhatnagar, V. P. & Jakobsson, T. *Geophys. Res. Lett.* **5,** 180 (1978).
57. Chiu, Y. T. & Straus, J. M. in *Possible Relationships between Solar Activity and Meteorological Phenomena* 83 (NASA SP-366, Washington, DC, 1975).
58. Thorne, R. M. *Science* **195,** 287 (1977).
59. Markson, R. *Nature* **273,** 103 (1978).

Predictability of climate

C. E. Leith

National Center for Atmospheric Research, P.O. Box 3000, Boulder, Colorado 80307

A mathematical framework is presented for the study of the predictability of small changes in the climatic system. From a statistical definition of climate in terms of ensemble averages, a limit to predictability is found arising from random weather fluctuations. This limit, typically referred to as climate noise, is often able to obscure a possible climate signal produced by external forcing. A separation of timescales of climatic response is also presented as a convenient means to distinguish between the fast internal (for example, the atmosphere) and the slow external (for example, the oceans) systems. In this way, it is possible to estimate the climate signal in the internal system that arises from small changes in the slow external system. Finally, a theoretical approach is presented for the inclusion of possible feedback effects in the estimation of a climate signal. Such feedback effects could result from the external climatic system being influenced by a change in the mean properties of the fast internal climatic system.

In recent years a number of widely publicised anomalies in various parts of the world have greatly increased interest in the predictability of climate for periods of months to decades. In this review I shall define climate, as distinguished from weather, in terms precise enough to serve as a basis for a statistical dynamical theory of small fluctuations. The problem is to separate potentially predictable fluctuations called signals from unpredictable fluctuations called noise. The ratio of signal to noise provides a measure of the predictability of the climate system.

Definition of climate

The climate at any place is commonly defined as being the average weather there. A broader definition of the global climatic system includes all the statistical properties of the global atmosphere and the underlying land, sea and ice surfaces. The climate system is thus described in terms not only of mean values but also of variances about the mean, co-variances between variables possibly separated in space and time, probabilities of extreme events, and so on. Specification of the statistical properties of the climatic system involves far more information than does specification of the state of the atmosphere at a given time, but little is known about the statistical moments of the system higher than the second.

The statistical properties of the climate system must be defined in terms of some sort of averaging process and here fundamental difficulties arise. The common, operationally feasible, and natural choice is to compute an average over some time interval T. This permits a continuum of choices depending on T. For small enough values of T the corresponding time-averaged climate would become indistinguishable in its fluctuations from the weather. For greater and greater values of T the corresponding time-averaged climate becomes increasingly

0028–0836/78/0276—0352$01.00

© Macmillan Journals Ltd 1978

independent of the averaging interval, and we might define a true climate in terms of a limit as T approaches infinity. In addition to the practical difficulty that computing an infinite time average is no longer operationally feasible, there is the major theoretical difficulty that we would now no longer be able to consider climate as changing with time. A conventional compromise is to take T as 30 yr.

Such a compromise definition of climate is awkward for setting up a precise mathematical framework for a theory of climate. This same general difficulty was encountered over a century ago in considering the formulation of a statistical mechanical basis for thermodynamics. In this analogy the detailed motions of gas molecules correspond to the weather and the statistical properties of a gas such as temperature and pressure correspond to the climate. Here too, if one wishes to discuss a situation in which temperature and pressure are changing, one must not define them in terms of infinite time averages.

Waterspout photographed near Florida Keys, 10 September 1969, by J. H. Golden of ERL-NOAA, Boulder, Colorado (note the spiral forming on the surface of the water)

In statistical mechanics this difficulty has been circumvented by introducing the theoretical concept of an ensemble. An ensemble is an essentially infinite collection of copies of the original dynamical system with each evolving in its own natural way. The dynamical state for each copy is specified by giving the values of a large number of dynamical variables, and it can therefore be represented by a point in a multi-dimensional dynamical phase space. At any given time the members of the ensemble will be found at phase points throughout the whole phase space in accordance with some probability distribution. In time the phase points move and thus in general the probability distribution changes. The problem in statistical mechanics is to find an equilibrium probability distribution that does not change with the natural motion of the phase points in order to define statistical properties in terms of the corresponding ensemble average. One may then ask sensible questions about how external influences can modify statistical properties by changing the equilibrium ensemble probability distribution.

The application of this general concept to the climate system is straightforward. One imagines an ensemble of Earths, each subjected to the same external influences, each with the same distribution of continents and oceans, and the weather on each going through normal fluctuations independently of all others. One next imagines that the distribution of states of the weather in the ensemble is chosen in such a way that the probability distribution is unaffected by the rapid weather fluctuations of individual members but can change relatively slowly owing to changing external influences. The climate is then defined in terms of averages over this ensemble. It is immediately evident that this ensemble-averaged climate is an ideal theoretical concept inaccessible by direct observation just as was the infinite time-averaged climate discussed previously. For a system with unchanging external influences, it is desirable for the two definitions to be equivalent. A dynamical system is said to be ergodic or transitive if this is the case.

At this point the objection might be raised that the ensemble-averaged climate is not really what anyone is interested in. We live on only one Earth with a single sequence of weather events and are naturally more concerned about the future behaviour of that single system than in some average behaviour of an imagined ensemble. Accordingly, the most direct way to predict the climate would seem to be to predict the weather and then to average it in time. Unfortunately, we are currently unable to predict most of the variability of the weather beyond a few days. Furthermore, theoretical studies indicate that there exists a limit of a week or so in the range of useful prediction even for a perfect prediction model because of the unstable nature of the atmosphere and the inevitable inadequacies in observations of its detailed initial state.

We adopt then the point of view that the climate is determined by an ideal ensemble and that finite time averages will fluctuate about ensemble values producing what is called climate noise. This climate noise, to be discussed in the next section, arises from the statistical fluctuations of the unpredictable weather and not only obscures the properties of the true climate but also diminishes the value of knowledge of these properties for prediction of the short-term behaviour of the most interesting member of the ensemble, the one which actually exists.

A further question arises in seeking a definition of climate precise enough for research purposes. How does one bound the statistical climate system to distinguish its internal behaviour from external forcing influences? Ideally, this should be done in such a way that the external forcing influences are independent of the state of the internal system. They may be taken, for example, to be the positions of the continents and the incoming solar radiation at the top of the atmosphere. For many model studies of the climate, however, it is useful to specify as external a much larger number of influences such as carbon dioxide content, sea surface temperature, ice sheet thickness and surface albedo. Although the separation between the internal and external systems may be made in many ways, it is important to remember that any feedback influences of the internal system on the external system must either be ignored or treated in an ensemble-average sense. Otherwise, the external system acquires some of the unpredictable random behaviour of the internal system and can no longer be sharply specified.

By judicious location of the boundary between the internal and external systems, we can also implicitly separate time scales of change. The advantages sought in an explicit choice of averaging time scale can instead be obtained by choosing the internal system to include all relatively fast processes. To achieve a time scale separation at a few weeks, for example, the internal system would contain the atmospheric motions and changing land surface temperature, whereas the sea surface temperature would be treated as an external forcing influence.

Climate noise

Perhaps the most familiar example of a sampling fluctuation occurs with the tossing of a coin. Although for a large number of tosses the ratio of heads to tosses is expected to approach one-half, no-one is greatly surprised if 20 tosses result in 7 heads and 13 tails. If 200 tosses, however, were to result in 70 heads and 130 tails, some doubt might be raised as to the fairness of the coin. The theory of such fluctuations is completely understood and can be found in any elementary textbook on probability theory.

A similar example more relevant to the continuous variables of the climate system is the drawing of a sample of n independent values x from an infinite population with mean μ and standard deviation σ. The average $\bar{x}$ of the sample generally will

differ from μ, and it is easy to show that many repetitions of the sampling procedure will lead to a distribution of $\bar{x}$ with mean μ and standard deviation $\sigma_n = \sigma/n^{1/2}$.

Averages of a meteorological time series over an interval T have similar fluctuations about the infinite population mean, but in this case sequential observations, if frequent enough, are not mutually independent. This may easily be accounted for[1,2] in terms of an effective time between independent samples

$$T_o = \int_{-\infty}^{\infty} R(\tau)\, d\tau$$

where $R(\tau)$ is the correlation of two values in the time series separated by a time interval τ. For $T/T_0 = n$ sufficiently large, n plays the role of the independent sample size. The standard deviation σ_T of time averages is then related to the standard deviation σ of the time series by

$$\sigma_T = \sigma/n^{1/2} = \sigma(T_0/T)^{1/2}$$

The time T_0, which is an integral time scale for fluctuations of meteorological variables, plays an important part in the determination of climate noise levels. Madden[3] has determined T_0 for sealevel pressure over much of the Northern Hemisphere and finds values with a range 2–8 d.

Tornadoes over Vietnam photographed from a helicopter. Taken from *The Elements Rage* by F. W. Lane (David and Charles, Newton Abbot, UK)

Some part of the fluctuations from year to year in monthly averages of meteorological variables must be attributed to the climate noise arising from sampling fluctuations. As such it represents an unpredictable component, and a key question for climate predictability is how much greater is the variability observed to be than the amount expected from noise. Madden[3] has attempted to answer this question for monthly average sea surface pressures for January and July. He has found that most of the observed interannual variability in midlatitudes seems to be climate noise, whereas in the tropics there is evidence of an additional potentially predictable component.

The existence of an unpredictable noise component in finite time averages limits the practical utility of any predicted shift $\Delta\mu$ in the climate mean μ which we may appropriately call a climate signal. Unless the signal-to-noise ratio $\Delta\mu/\sigma_T$ is greater than about 0.5, the prediction of $\Delta\mu$ may have little utility for prediction of time averages of duration T (ref. 1). Of course, as σ_T decreases at $T^{-1/2}$, for long enough intervals, T, the prediction of $\Delta\mu$ acquires value.

Climate signal

In the previous section I described the climate noise that arises from unpredictable weather fluctuations in the fast internal climatic system. In this section I shall discuss procedures for estimating the climate signal that arises from the influence of the slow external system on the internal system. To the extent that the slow system may be predictable, the coupling of the two systems provides an element of predictability for the fast system.

The usual procedure for determining the response of the internal system is to construct a numerical model of it in which the external influences are treated as specified forcing terms. Symbolically, we may write the model dynamical equations for the unperturbed system as

$$\dot{x}_\alpha = Q_\alpha(x_1, x_2, \ldots, x_N) + f_\alpha$$

Here $x_1, x_2, \ldots, x_N$ are amplitudes of a large but finite number N of modes in terms of which the state of the internal system is defined, $Q_\alpha(x_1, x_2, \ldots, x_N)$ are complicated functions determining the internal dynamics of the system and f_α are forcing terms describing the external influences. For the statistically stationary unperturbed system, we have unchanging ensemble means, that is, $\langle \dot{x}_\alpha \rangle = 0$. If now a step change in forcing δf_α is introduced at time $t = 0$ for all members of the ensemble, then the ensemble means will start to change at a rate $\langle \dot{x}_\alpha \rangle = \langle \delta f_\alpha \rangle$. Because of the stable nature of the climatic system, a new stationary level will be approached with a final change in means given by $\delta\langle x_\alpha \rangle$.

The general procedure in using a numerical model to estimate $\delta\langle x_\alpha \rangle$ is to make long computations with and without the perturbed forcing δf_α, to estimate $\langle x_\alpha \rangle$ in either case by time averages over intervals excluding any initial transient, and to subtract one estimate from the other. The climate noise associated with the time averages tends to obscure small values of $\delta\langle x_\alpha \rangle$. It is usually assumed that small responses are linear and that δf_α may be amplified by a scale factor by which the computed response is later divided.

The first-order linear response of the system can be written as

$$\delta\langle x_\alpha \rangle = \sum_\beta A_{\alpha\beta}\, \delta f_\beta$$

where the matrix elements $A_{\alpha\beta}$, having the dimensions of time, determine the response of the mean of any mode $\langle x_\alpha \rangle$ to a change in forcing of any mode δf_β.

Another procedure provides more detailed information about the time history of the response. Consider an impulsive forcing which is turned off immediately after being imposed at time $t = 0$ on each member of the ensemble. This produces an immediate change $\delta\langle x_\alpha(0^+) \rangle$ which will then tend to zero as the system returns to its original statistical state. For small perturbations the relaxation is described by

$$\delta\langle x_\alpha(\tau) \rangle = \sum_\beta g_{\alpha\beta}(\tau)\, \delta\langle x_\beta(0^+) \rangle$$

where $g_{\alpha\beta}(\tau)$ is a mean linear response matrix function of time τ with $g_{\alpha\beta}(0^+) = \delta_{\alpha\beta}$ the identity matrix. The response to any history of forcing perturbation $\delta f_\alpha(t)$ may be written in terms of $g_{\alpha\beta}(\tau)$ as

$$\delta\langle x_\alpha(t) \rangle = \sum_\beta \int_{-\infty}^{t} g_{\alpha\beta}(t-s)\, \delta f_\beta(s)\, ds$$

In particular, for step forcing the infinite time response matrix is given by

$$A_{\alpha\beta} = \int_0^\infty g_{\alpha\beta}(t)\, dt$$

A determination of $g_{\alpha\beta}(\tau)$ from model calculations is probably not feasible, but it can be estimated by statistical analysis of the unperturbed climate system itself. In the course of their natural evolution, the members of the ensemble generate anomalies δx_α from the mean. For a sub-ensemble all of whose members have a given anomaly $\delta x_\alpha(0)$ at time $t = 0$, the sub-ensemble mean perturbation will be $\delta\langle x_\alpha(0) \rangle = \delta x_\alpha(0)$. But it, too, will tend to return to zero in a way that can be estimated by linear regression prediction methods.

The timelagged co-variance matrix for the unperturbed system is

$$X_{\alpha\beta}(\tau) = \langle \delta x_\alpha(t+\tau)\, \delta x_\beta(t) \rangle$$

the regression matrix is

$$R_{\alpha\beta}(\tau) = \sum_\gamma X_{\alpha\gamma}(\tau)\, [X^{-1}(0)]_{\gamma\beta}$$

where $R_{\alpha\beta}(0) = \delta_{\alpha\beta}$, the identity matrix, and the least square estimate of the future mean anomaly of the sub-ensemble is given by

$$\delta\langle x_\alpha(\tau)\rangle = \sum_\beta R_{\alpha\beta}(\tau)\langle \delta x_\beta(0)\rangle$$

The reasonable assumption that the system recovers from a small natural anomaly in the same way that it does from one induced by external forcing leads to the fluctuation dissipation relationship that for $\tau \geqslant 0$

$$g_{\alpha\beta}(\tau) = R_{\alpha\beta}(\tau)$$

According to the fluctuation dissipation theorem[4–7], this relationship holds exactly for dynamical systems with a gaussian Maxwell–Boltzmann equilibrium probability distribution. Such is not the case for a dissipative turbulent system such as the atmosphere, but experience with turbulence models has indicated that the approximation may nonetheless be fairly good.

One is faced with the choice of estimating $g_{\alpha\beta}(\tau)$ for the internal climatic system either as given by $R_{\alpha\beta}(\tau)$ for the system itself or as given by $g_{\alpha\beta}(\tau)$ for a numerical model of it. Confidence in the results would be greatly enhanced if the two quite different approaches gave similar answers, and this confidence would extend to model results for perturbations beyond the linear domain of the fluctuation dissipation relationship.

Qualitative properties of climate signals may be deduced from known qualitative properties of correlations using the fluctuation dissipation relationship, as regression functions are qualitatively like correlation functions. These have a spatial extent in midlatitudes of a few thousand kilometres, and the climate response to a local forcing perturbation should be similarly limited in spatial extent. This suggests that the climatic system may be considered as a collection of more or less independent regional climatic systems.

As the diagonal regression coefficients $R_{\alpha\alpha}(\tau)$ are qualitatively like the single variable lagged correlation function $R(\tau)$ discussed in the previous section, a qualitative estimate of the step response coefficients for the fast internal system is given by

$$A_{\alpha\alpha} = \int_0^\infty R_{\alpha\alpha}(\tau)\, \mathrm{d}\tau \approx T_0/2 \approx 3 \text{ days}$$

The internal climatic system adjusts therefore within a week or so to a new level with a change in mean values $\delta\langle x_\alpha\rangle$ given by the initial rate of change δf_α multiplied by a characteristic response time of a few days.

Feedback effects

It has been assumed that the external climatic system is not influenced by the mean or fluctuating properties of the fast internal climatic system. The feedback influence of mean effects, however, may be included rather easily to first order by linear analysis. It is convenient here to replace matrix elements by matrix notation so that the step forcing response equation becomes

$$\delta\langle \mathbf{x}\rangle = \mathbf{A}\, \delta\mathbf{f}$$

Suppose now that a change in mean properties $\delta\langle\mathbf{x}\rangle$ in the internal system forces a change in the external system leading eventually to an additional change in forcing $\delta_1\mathbf{f}$ which to first order is given by

$$\delta_1\mathbf{f} = \mathbf{B}\, \delta\langle\mathbf{x}\rangle$$

The combined effects are then given by

$$\delta\langle\mathbf{x}\rangle = \mathbf{A}(\delta\mathbf{f} + \mathbf{B}\delta\langle\mathbf{x}\rangle)$$

so that we have

$$\delta\langle\mathbf{x}\rangle = (\mathbf{I}-\mathbf{B})^{-1}\, \mathbf{A}\, \delta\mathbf{f}$$

The feedback effect matrix $(\mathbf{I}-\mathbf{B})^{-1}$ can either increase or decrease the response $\delta\langle\mathbf{x}\rangle$ from that with $\mathbf{B}=0$.

The response of the external system is assumed to be slow and, for this present analysis, non-random. The response properties leading to the matrix $\mathbf{B}$ may therefore be determined relatively easily using a numerical model of the external system that includes the appropriate slow physical processes.

As with the internal system, it is possible to determine the more detailed time history of the feedback response of the external system. Qualitatively, a step change $\delta\mathbf{f}$ induces a rapid change $\delta_0\langle\mathbf{x}\rangle$ which induces a slow change $\delta_1\mathbf{f}$ which induces a further slow change $\delta_1\langle\mathbf{x}\rangle$ to the final value $\delta\langle\mathbf{x}\rangle = \delta_0\langle\mathbf{x}\rangle + \delta_1\langle\mathbf{x}\rangle$ given above.

The additional feedback influence of the random fluctuations in the internal system has been discussed by Hasselmann[8] and others[9,10]. Now the external system becomes random and Hasselmann considers one that can be modelled by a Langevin equation of the general form

$$\dot{\mathbf{y}} = -\mathbf{D}\mathbf{y} + \mathbf{g}$$

Here $\mathbf{y}$ represents anomalies in the external climatic system, $\mathbf{D}$ is a matrix which is positive definite to provide stability and $\mathbf{g}$ is a random white forcing term that models the effects of the rapid fluctuations of the internal system. The resulting behaviour is analogous to damped brownian motion with the external system playing the part of the particle and the internal system that of the molecular motions. By suitable choice of $\mathbf{D}$ and the strength of $\mathbf{g}$, it has been possible to model the frequency spectrum of sea surface temperature fluctuations considered as part of the external system[9,10]. Some predictability of the external system is provided for the Langevin models by the mean regression equation

$$\langle\dot{\mathbf{y}}\rangle = -\mathbf{D}\langle\mathbf{y}\rangle$$

Discussion

In this review I have described the mathematical framework for the study of the predictability of those small changes in the climate for which a linear analysis is appropriate. For larger changes a non-linear numerical climate model must be used.

Much work has yet to be done to fill in this framework. There is already fair evidence for useful climate predictability in equatorial and tropical regions where the noise level is small, but there is only fragmentary evidence for it in midlatitudes.

Finally, the possibility of extending the range of deterministic prediction of relatively stable internal configurations of the sort that arise in atmospheric blocking events is of great practical importance. But this would be a prediction of weather rather than of climate as strictly defined in this review.

The National Center for Atmospheric Research is sponsored by the NSF.

1. Leith, C. E. *J. appl. Meteor.* **12,** 1066–1069 (1973).
2. Jones, R. H. *J. appl. Meteor.* **14,** 159–163 (1975).
3. Madden, R A. *Mon. Weath. Rev.* **104,** 942–952 (1976).
4. Callen, H. B. & Greene, R. F. *Phys. Rev.* **86,** 702–710 (1952).
5. Kraichnan, R. H. *Phys. Rev.* **109,** 1407–1422 (1958).
6. Kraichnan, R. H. *Phys. Rev.* **113,** 1181–1182 (1959).
7. Leith, C. E. *J. atmos. Sci.* **32,** 2022–2026 (1975).
8. Hasselmann, K. *Tellus* **28,** 473–485 (1976).
9. Frankignoul, C. & Hasselmann, K. *Tellus* **29,** 289–305 (1977).
10. Reynolds, R. W. *Tellus* **30,** 97–103 (1978).

Predicting temperature trend in the Northern Hemisphere to the year 2000

M. K. Miles*

The cooling of the Northern Hemisphere since 1940 has been variously interpreted as the overture to the next Ice Age, the effect of industrial pollution in the atmosphere or of a decline in the solar output. Are we in a position to judge between these various interpretations and to make a prediction for the next few decades? The link with the next Ice Age may be dismissed as a confusion of timescales; the explanation in terms of atmospheric pollution merits careful examination but seems unlikely to be adequate on its own. Natural fluctuations must also be considered.

THE annual surface temperatures of the Northern Hemisphere from 1951 to 1975 given by Kukla *et al.*[1] show that year to year differences equalled or exceeded 0.3 °C eight times in this period. These changes do not constitute climate as generally understood, although they may be very important for world agriculture or transportation. There are various views on how long a period we should average this temperature over to arrive at the value which is to be the subject of climatic predictions.

How 'noisy' is the temperature record?

The data themselves do not give us much guidance: variations seem to occur on all time scales from a year to millennia. Claims have been made at various times that at some particular frequency (2.2 yr, 11 yr, 88 yr, 200 yr, and so on) there is a marked periodicity, but the variability associated with any of these is usually a small part of the total: no period stands out so clearly from the rest that we can on this basis select an appropriate averaging period.

The spectrum is not, however, simply a 'white noise' one with all frequencies equally represented: there is a notable tendency for there to be more variation at periods above about 10 or 20 years than a white noise spectrum would contain. This I interpret as meaning that it may not be a fruitless task to try and say something about the future trend of means taken over 5 or, perhaps better, 10 yr.

Figure 1 shows the course of decadal means of Northern Hemisphere surface temperatures for the past 100 yr. These have a standard deviation of 0.17 °C but the deviations are not randomly distributed. A random series with this standard deviation would show average decade to decade differences of about 0.24 °C, whereas this series has an average difference of 0.11 °C between adjacent decades. Moreover, since 1915 there have been three adjacent positive values (that is, a warming trend) followed by three successive negative values—a pattern which hints strongly at a dominance of non-random effects.

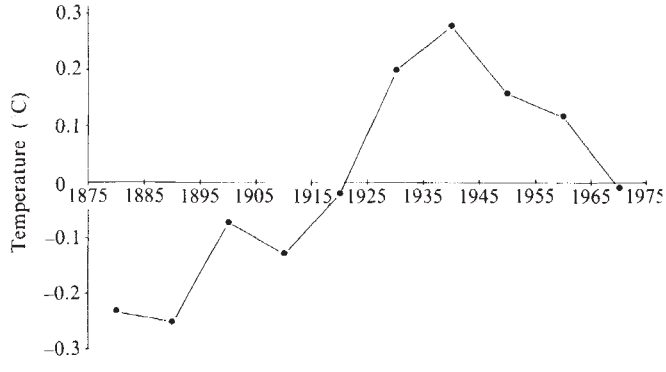
If we suppose for the moment that a warming trend of 0.1 °C per decade from 1885–1944 followed by a cooling trend of the same magnitude to 1974 represent non-random effects, then the decadal values have a random fluctuation corresponding to a standard deviation of 0·05 °C superimposed on this pattern. This is certainly a minimum value of the 'noise', compounded of instrumental errors and natural fluctuations not taken into account. As most explanations do not account for more than about two-thirds of the variance over the past 100 yr, a better working estimate for the standard deviation of the noise is probably 0.1 °C. This means that there is a 5% chance that the next decade (1975–84) will be ±0.2 °C or more different from 1965–74 due to 'noise'. Thus, a continuance of, or reversal of, the trend could not reliably be inferred from such a change. If, however, the following decade showed a further 0.1 °C or more movement in the same direction this could be considered as an organised trend with a high probability that it was not due to chance.

From such considerations one could set up three reasonable categories for predictions of the decade centred around AD 2000: (1) 0.3 °C or more above the value for 1965–74, implying a reversal of the cooling trend of the past 30 yr with temperature at least back to that of the optimum in Fig. 1; (2) 0.3 °C or more below the value for 1965–74, implying a continuance of the cooling trend at about the same rate; and (3) within ±0.3 °C of 1965–74 which implies an expectation of random fluctuations about the value of the latest decade.

Some predictions or quasi-predictions

The published predictions are not always quantified; they tend to say either that the cooling trend will continue or that it will reverse. Among those favouring continued cooling, Lamb has sometimes mentioned a return to the conditions prevailing at the end of the eighteenth century, which would put his forecast in category (2); and Bryson[3] is such an enthusiastic advocate of cooling that his can hardly be put in category (3). Budyko and Vinnikov[4], the most emphatic advocates of a warming, predicted a value for AD 2000 of between 0.5 and 1.0 °C above the latest decade—a prediction clearly in category (1). Mitchell[5] expected that the warming effect of CO_2 and the cooling effect of man-made dust were likely to be in near balance to the end of the century. One might deduce from this that natural fluctuations would determine the sign of the deviation. His expectation could probably be assigned to category (3), but he considers a specific prediction to be unwarranted. Kellogg[6] likewise thinks that we are not in a position to make a categorical forecast: he considers that a warming of 0.5–1.0 °C due to increasing CO_2 is fairly likely by AD 2000, whereas cooling due to man-made aerosol is more doubtful and is likely to be less than this. Thus, he concludes that man's activity will probably raise Northern Hemisphere temperature by some 0.5 °C by the end of the century and natural fluctuations are unlikely to mask this warming trend.

Fig. 1 Decadal means of surface temperature for the Northern Hemisphere, plotted as deviations from the average, 1875–1974. (After Budyko, Willett and Mitchell to 1969; updated to 1974 from Kukla *et al.*[1].)



*Formerly of The Meteorological Office, London Road, Bracknell, Berks, UK

0028–0836/78/0276—0356$01.00

© Macmillan Journals Ltd 1978

Rationale of the predictions

Budyko and Vinnikov[4] and Bryson[3] interpret the temperature trends over the past 100 yr in terms of the changing transmissivity of the atmosphere. They base this on a Russian time series of direct radiation measurements first assembled by Pivovarova in 1968 and later updated by Budyko and Vinnikov[7]. They are shown in Fig. 2 as deviations in percentage from the average for 1885–1970. The data for the early decades come mainly from Russian observatories but after about 1914 data from American observatories are included. There is an increase of about 4% in direct radiation between the turn of the century and 1940 followed by a decrease of about 6% by the late 1960s. Budyko and Vinnikov[4] and Bryson[3] attribute the increase mainly to the clearing of the atmosphere following the series of violent volcanic eruptions in the 1880s, and the decrease after 1940 to the increased concentration of man-made dust due to industrial activity. From observations of the ratio of scattered to direct radiation at the surface Budyko and Vinnikov[7] conclude that the 4% gain of direct radiation implies a net gain of something under 1% (there is less scattered radiation) which is more than enough to explain the 0.6 °C rise in the Northern Hemisphere temperature over that period.

On the other hand there is considerable doubt whether enough of the volcanic dust ejected into the stratosphere by the Alaskan eruption in 1912 (Katmai)—the last major eruption until the 1950s—would be still present even 10 yr later to depress the temperature noticeably below a clear-air value. Yet about half of the increase in temperature occurred after that date, and Miles and Gildersleeves[8] have argued that this cannot be attributed to the clearing of volcanic dust. Although there is considerable evidence that dust after major eruptions depresses the hemisphere temperature by 0.2–0.3 °C for about 2 yr afterwards, the effect of a single major eruption on the decadal temperature is not likely to exceed 0.15 °C.

The series of major eruptions around the time of Krakatoa (1883) may have depressed the temperature of the 1880s by up to 0.3 °C but after that there has not been a comparable series of eruptions. The Russian series of direct radiation measurements shown in Fig. 2 suggests that the radiation recovered to a normal level after the eruptions of the 1880s, after those of 1902–03 and the Alaskan eruption in 1912. Although there is still some doubt about the response time of the atmosphere, it is likely that the hemispheric temperature returned to near the clear-atmosphere value after each. Thus the depressive effect of these volcanic episodes cannot be considered additive. If this is accepted, then the magnitude of the warming due to the clearance of volcanic dust can hardly be placed higher than 0.3 °C. This leaves most of the warming after 1920 to be explained in some other way and also the 1% increase in direct radiation after 1920. It is probable, however, that the radiation series is not homogeneous, as the number of stations used was increasing with time and it is difficult to be sure that calibrations are the same within 1% over these long periods.

The interpretation of the 6% decrease after 1940 presents further difficulties. First, it has not been demonstrated beyond a doubt[9] that the average atmospheric dust loading of the Northern Hemisphere has increased. The transmissivity at some locations has not changed during this century. It seems very probable that increased industrial activity has put more dust into the lower atmosphere but it seems that the bulk of it falls out within a few hundred kilometres of the entry zone, so that the average increase in tropospheric dust over the whole hemisphere is a very small amount indeed. The reduction in direct radiation shown by the Russian series, if due to man-made dust, is therefore considered to apply only in the neighbourhood of the observatories in Eurasia and North America.

Second, dust absorbs solar radiation as well as scattering it, so that there is a warming effect arising from the absorption of the direct solar radiation. In the case of volcanic dust (in the stratosphere) the warming is confined to the stratosphere but if the bulk of the additional man-made dust is fairly low down in

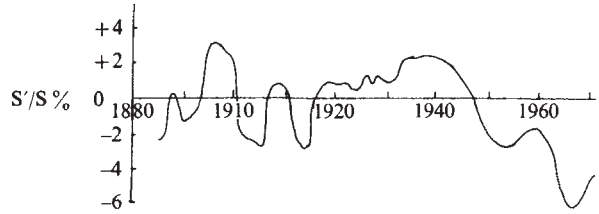


Fig. 2 Five-year means of direct solar radiation anomalies, (From Budyko and Vinnikov[7], based on work by Pivovarova.)

the troposphere this warming could affect the surface temperature. It is uncertain how much this will offset the loss by scattering. Budyko[4] reduces the cooling by one-third and, after making a further allowance for the fact that much of this dust will be below cloud level, concludes that this 6% reduction in direct radiation will cause a temperature reduction of 0.5 °C. This together with a warming of about 0.2 °C—a widely accepted estimate—due to the additional CO_2 gives a net cooling of the right amount. Bryson[3] considers the two effects together and arrives at a net cooling of 0.3 °C. Although the two estimates are in agreement, the uncertainties in reaching them are so great that it is important to explore other ways of assessing whether they are likely.

If the dust absorbs long-wave terrestrial radiation giving a so-called greenhouse effect then the net cooling due to the scattering effect may be very small indeed. Indeed Idso and Brazel[10] have put forward evidence that with realistic concentrations of dust the net effect on the surface temperature would be a warming. Without going this far it seems likely that the surface cooling due to a given concentration of man-made dust in the lower troposphere will be less than that due to the same concentration of volcanic dust in the stratosphere. Mitchell[5] has given estimates of the loading of man-made dust which show that at the end of the 1960s it was about half the estimated average loading of stratospheric volcanic dust in the 1880s. If the depression of Northern Hemisphere temperature below a clear-air equilibrium value for the 1880s could be reliably established, this would give an upper limit to the cooling to be expected from the recent concentration of man-made dust.

Already a divergence of opinion appears. Bryson[3] (and to some extent Budyko[4]), in saying that practically the whole of the 0.6 °C warming from the 1880s to the 1940s is due to the clearing of the volcanic dust, is putting this upper limit at 0.6 °C. Lamb[2] is inclined to the view that other factors were as important in bringing about the warming. It is difficult from the hemispheric temperature data to see how the depressive effect in the 1880s could have exceeded about 0.3 °C and if the effects of the successive eruptions are not additive this would be the upper limit to the cooling effect of the man-made dust in the late 1960s.

Agreeing as they do on the cause of the post-1940 cooling, how is it that Bryson and Budyko then reach such divergent views about the future trend? Bryson regards the cooling since 1940 as showing that the extra opacity due to industrial activity has a larger effect on surface temperature than the reduced emissivity due to the increased CO_2 arising from the same industrial activity. Thus, the expected increase in industrial activity during the rest of this century will bring a steadily decreasing surface temperature. Budyko, however, says that present pollution control measures are reducing the input of dust but not that of CO_2 and that eventually the warming effect of CO_2 will be decisive. He advances two pieces of evidence that this is already happening. The amount of direct solar radiation has, he says, increased since the low values of the 1960s, and the hemispheric temperature has risen by 0.3 °C in the past decade. The reservations concerning the radiation measurements have already been discussed and no other climatologist has claimed to find such a marked increase in temperature for the Northern Hemisphere. Most estimates indicate a reduction of the cooling rate almost to zero in the 1970s but do not show a reversal.

Sorry, for copyright reasons some images on this page may not be available online

Campbell–Stokes sunshine recorder photographed by W. Scriven, Great Dunn Fell, Westmoreland (now Crown Copyright, Meteorological Office, Bracknell, UK)

Budyko supports his claim for a hemispheric warming by pointing to a warming of about 1.0 °C in the polar regions since the mid-1960s. This polar warming also emerges from an analysis of data from ice flow stations. It reduces the size of the meridional temperature gradient and he claims that in the past a reduction in meridional temperature gradient has been generally associated with an increase in the hemispheric temperature. There has however, been at least one period in the past (1935–44) when the polar temperature was decreasing while that of the hemisphere was increasing, and it would be advisable to have more direct evidence of a hemispheric warming before accepting Budyko's thesis.

One of the secondary reasons for believing that dust has led to such a substantial cooling since 1940 is that the observed cooling has to be explained in the face of an expectation that the increased CO_2 has increased the equilibrium temperature since the turn of the century by about 0.3 °C. This is widely regarded as the one effect known most precisely but it is not entirely satisfactory to use it as indirect support for the man-made dust case. The net effect of the CO_2 with cloud feedback may be substantially less than this. Besides, if the further increase in temperature after 1920 is neither due to the clearing of volcanic dust nor to the increased CO_2 then a return to the level of the 1920s, if the causative mechanism ceased after 1940, would not be altogether surprising.

There are moreover some grounds for thinking that 0.3 °C is an overestimate of the hemispheric cooling since 1940. Sea temperature data for the tropical Atlantic and Pacific now being analysed in the UK Meteorological Office indicate that temperature increased from a minimum in the 1920s until the early 1960s, and the mean value for the 1960s is probably higher than that for the 1930s. There is some reason to think that the published estimates of Northern Hemisphere temperature (on which the graph in Fig. 1 is based) gave insufficient weight to temperatures in the tropics. If the course of the surface air temperature of the tropics followed that of the sea temperature, then the cooling since 1940 would probably be less than 0.3 °C. This perhaps strengthens the man-made dust case because one of the objections against it is the magnitude of the cooling effect claimed. On the other hand, if pollution control measures are increasing the atmospheric transmissivity as Budyko (and some others) are claiming, then the warming due to increasing CO_2 may more easily balance the cooling effect of dust and even become the dominant effect before the end of the century.

Prediction of some natural fluctuations

If these two man-made effects on climate are likely to be in near balance until the end of the century, this would leave the outcome to be determined by natural fluctuations. These may be due to variations in solar output or planetary albedo (apart from changes due to man-made dust) on the one hand and differences in the amount of heat stored in the oceans arising perhaps from variations in the general wind circulation on the other. Mitchell[5] and Kellogg[6] believe that we are not in a position to predict any of these effects. Some other climatologists are not so cautious and make a prediction of continued cooling based on indirect evidence of a change in solar output associated with long-term variations in the amplitude of the 11-yr sunspot cycle. The number of dark spots on the Sun increases from near zero to a maximum and then back to near zero over a period of about 11 yr. The size of the maximum has varied in the past with a somewhat irregular periodicity of 80–90 yr and there are some grounds for thinking that there is a periodicity of about 180 yr (and perhaps an even longer one). Whether the heat output from the Sun varies with the amount of sunspot activity has not yet been determined definitively. Direct measurements of the solar constant from satellites are not yet adequate to give the answer. Attempts to obtain it from measurements within the atmosphere have likewise failed because corrections for atmospheric turbidity are as large or larger than the changes being looked for and cannot be made with sufficient accuracy.

The views that are held on this matter arise from attempts to relate hemispheric temperature directly to the sunspot number. There seems to be a positive relationship whereby temperature is higher when the sunspot activity is high, but it is a weak relationship with such a low statistical significance that it could

Sorry, for copyright reasons some images on this page may not be available online

Thermopile for measuring solar radiation (Crown Copyright, Meteorological Office, Bracknell, UK)

be a coincidence. As the reliable data on hemisphere temperature extend over little more than one cycle of the 80–90 yr periodicity, this contributes to the uncertainty about the result. There are, moreover, differences in the apparent relationship from epoch to epoch and this is especially notable when we search for the effect in the 300-yr-long temperature record for central England assembled by Manley.

Those who think such analyses nevertheless give a dependable result, that is, higher hemispheric temperatures at the peak of the 80–90 yr cycle, have drawn further support from the analysis of an ice core taken from the Greenland Ice Cap at Camp Century in 1966. This showed prominent cycles at 77 and 177 yr which have been interpreted as being forced by the sunspot variations. The value of this as independent evidence is weakened by the difficulty of constructing the depth/time scale for the ice core and the admission that in this construction an effort was made to preserve natural climatic fluctuations of these lengths, that is, by assuming what we are trying to prove.

The sunspot activity in fact increased with each cycle from 1900, reached a high value in the cycle 1954–64 and fell considerably in the next cycle. It seems a reasonable extrapolation of past solar behaviour to expect the remaining two cycles of this century to show further reductions, and on the basis of the above relationship a number of forecasts of a cooling trend have been made. The proponents of these views rarely state what amplitude the Sun-linked cycles may have. The optimum fit of central England temperature data for spring, summer and autumn—the seasons with the best correlation between temperature and sunspot numbers—for a cycle of 88 yr (that is, eight 11-yr sunspot cycles) gives an amplitude of 0.13 °C which is barely above the noise for this data. Accepting this and placing the maximum in the 1954–64 cycle we should expect the coolest 11-yr period (cycle 1997–2007) to have a temperature 0.26 °C below that of 1954–64. This is for central England data which has in the past displayed variations with nearly twice the amplitude of the Northern Hemisphere data. The amount of depression of hemisphere temperature at the minimum of this cycle around AD 2000 would seem to be at most 0.15 °C. This is just about the amount of cooling between the decades 1955–64 and 1965–74, and would not in any case justify an unambiguous forecast in category (2), that is, a continuance of the cooling trend.

What do we know of the effect of changes in the large-scale circulation on the cloud cover or the distribution of heat between ocean and atmosphere? At present we neither have the knowledge nor the data to know precisely what these changes have been in the past. The intensity of the circulation seems to have been significantly higher during the warming period (1900–40) than in the subsequent years of decreasing temperature. The association of the warming of the Russian Arctic with the increased circulation on the North Atlantic also points to a positive association of circulation strength with hemisphere temperature. The correlation coefficient of 5-yr means of the surface westerlies of middle latitude and hemisphere temperature is, however, not above the chance level. The low correlation arises because the former peaked in the decade 1920–29 and the latter at least 20 yr later. The correlation with the temperature lagging 20 yr is statistically significant but it is not easy to find a satisfactory physical explanation for a lag of this amount. Even if we accept that there is a real association between circulation strength and temperature we need to be able to predict the former to exploit it in a temperature forecast. There is no adequate explanation of the high level of the circulation in the period 1900–40. Some Russian climatologists, among them Vitels (see Lamb[2]), have claimed that the strength of the mid-latitude westerlies is positively related to the long-period sunspot activity. The correlations, however, fall in the range where they could be coincidence especially as the circulation data cover only one cycle.

In an attempt to overcome this difficulty Lamb[2] has assembled data on the frequency of surface south-westerly winds at London going back to 1340 and, assuming these to parallel the strength of the general circulation, has proposed that there is a 200 yr cycle in its strength. On this basis the decline in circulation since 1940 may be expected to continue to the end of this century and beyond. Supposing this to be correct a positive relationship with temperature would then indicate a continued cooling.

Accepting for the moment that the series of south-westerly wind frequencies is homogeneous over the three centuries, there is still the need to show that their frequency is a good measure of the strength of the general circulation. Lamb[2] has not established this beyond a doubt but there looks to be a strong likelihood that this is the case. The next step, that is, establishing a 200-yr cycle, is done less satisfactorily. The argument from the surface wind data seems to rest on the occurrence of times of rapid decline in frequency at intervals of 200 yr from the middle of the sixteenth century to about 1960: this would perhaps establish a saw-toothed highly asymmetrically profile but would seem to leave an uncertainty about further detail in the bulk of each 200-yr period. He adduces evidence of a 200-yr oscillation in solar output as a forcing factor for this putative cycle in the circulation intensity. This evidence comes from variations in ^{14}C concentration thought to be due to variations in the corpuscular and far ultraviolet part of the solar output, the effect of which on the tropospheric circulation has not been satisfactorily established.

Summary of situation

The man-made changes in CO_2 concentration and dust seem to offer the clearest indication of the future trend of temperature. However, the need to explain a cooling over the past three decades in the face of an expected warming due to the additional CO_2 has tended to lead to an exaggeration of the cooling effect of volcanic dust and perhaps even more that of man-made dust. There is also a considerable doubt about how much extra optically effective dust there is on a global basis.

Inasmuch as there is still some doubt about the cause of the bulk of the warming after 1920, a considerable element of uncertainty is introduced into prediction. If the warming was a complicated end-effect of the enhanced circulation during 1900–40 resulting in a reduction in the area of Arctic pack ice and a consequently higher equilibrium temperature due to a reduced albedo, then the post-1940 cooling may be an effect of the reduced circulation. Up to the 1970s, however, it has not been accompanied by a return of Arctic ice to the higher values prevailing at the turn of the century, so that the equilibrium temperature arising from the ice-albedo contribution is probably still well above that around 1900. Does this mean that the full effect of the reduced circulation has not yet been reached? In this connection the failure to explain the lag of temperature behind the circulation is yet another example of how failure to explain the past temperature changes renders prediction hazardous.

In the absence of an explanation based on a consistent theory of climate, prediction would only be reliable if there were some strong and dominant periodic effects. Despite years of search for these and many confident claims of success, none of them stands up well to critical scrutiny. In these circumstances therefore only a great external pressure to make a categorical forecast or an undue attachment to one set of data and its interpretation would seem to explain the abandonment of an agnostic position, or at most the expression of a very cautious probability forecast.

1. Kukla, G. J. *et al. Nature* **270**, 573–580 (1977).
2. Lamb, H. H. *Climate: Present, Past and Future*, Vol. **I**, Fig. 6.8 and Vol. **II**, page 701 (Methuen, London, 1972 and 1977).
3. Bryson, R. A. & Dittberner, G. J. *J. atmos. Sci.* **33**, 2094 (1976).
4. Budyko, M. I. & Vinnikov, K. Ya. *Met. i Gidr.* **7**, 16–26 (1976).
5. Mitchell, J. M. in *The Changing Global Environment* (ed. Singer, S. F.) 164 (Reidel, Dordrecht, 1975).
6. Kellogg, W. W. in *Climatic Change* (ed. Gribbin, J.) 205–227 (Cambridge University Press, 1978).
7. Budyko, M. I. & Vinnikov, K. Ya. *Met. Hydr.* **9**, 3–13 (1973).
8. Miles, M. K. & Gildersleeves, P. B. *Nature* **271**, (1978).
9. Ellsaesser, H. W. in *The Changing Global Environment* (ed. Singer, S. F.) 235–272 (Reidel, Dordrecht, 1975).
10. Idso, S. B. & Brazel, A. J. *Science* **198**, (1977).

articles

Vertical particulate spires or walls within the Florida Current and near the Antilles Current

John R. Proni, Fred C. Newman, Feodor Ostapoff & Donald J. Walter

National Oceanic and Atmospheric Administration, Atlantic Oceanographic and Meteorological Laboratories, Ocean Acoustics, Miami, Florida 33149

Spires or humps of particulate material extending from the bottom of the oceanic mixed layer to near the ocean's surface have been observed in the Florida Current and in the regions of other current systems, using an ultrasensitive, 20-kHz acoustical echo sounding system. These spires may be the result of oceanic circulation rather than a purely biological phenomenon.

ACOUSTIC observations of spires of particulate matter extending from the bottom of the mixed layer to near the surface of the ocean have been made in the Florida Current (November 1977), in the region of the Antilles Current (January 1973) and in the Caribbean (April 1978). The Florida Current observations were made in the vicinity of 80° W long, and in latitudes between 25 and 28° N; Antilles Current region observations were made in around 70° W long, 25° N lat, and the Caribbean observations were made in a region ~200 nm north of Caracas, Venezuela. The observations were made using an ultrasensitive, short pulse length (<1 ms), towed, narrow beam (18°), 20 kHz acoustic echo sounding system. The use of this system to detect layers of particulate matter and internal waves in the ocean has previously been documented[1,2]. The acoustical frequency (20 kHz) and receiver sensitivity are such that scattering from oceanic turbulence or density gradients is not detectable, so that particulates and/or biota are the most likely sound scatterers detected.

Acoustic and XBT data

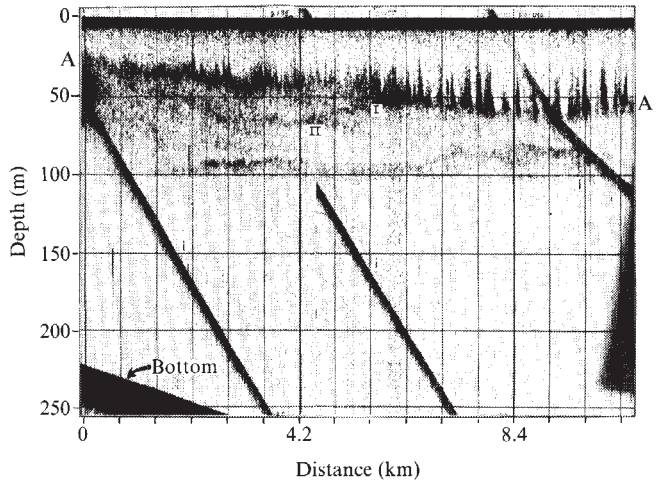
Figure 1 shows real-time acoustic data obtained from the RV Researcher while crossing the Gulf Stream off Palm Beach, Florida. The ship proceeded due east from Palm Beach, at a speed of 2 m s^{-1} beginning in water of depth 220 m. A layer of particulate matter corresponding to the depth of the bottom of the mixed layer may be seen gradually descending spatially from left to right (labelled A in Fig. 1). Lines of constant density (isopycnals) are known to slope down proceeding from west to east across the Gulf Stream[3]. The slope of the reflecting layer has a value $\sim 4\times10^{-3}$ which compares favourably with the 3×10^{-3} isopycnal slope, a value deducible from the isopycnal data of Schmitz and Richardson[3]. Interference lines generated by the Researcher's echo sounder are indicated and should be disregarded.

The most interesting features are the spires of particulate material which are seen to extend vertically from the particulate layer corresponding to the bottom of the mixed layer. Figure 1 shows that the spires eventually disappear in a shorewards direction; the mean horizontal spacing of the spires is ~200 m. Spire widths at the base of a spire seem to be 100 m or more. An important characteristic of the acoustical system used in gathering the data presented in Fig. 1, is that it was not designed for obtaining data near the ocean's surface, say, within 10 m. Accordingly, information or acoustical signals coming from this region is suppressed by the electronics. Thus it is uncertain whether the particulate spires observed in Fig. 1 reach the ocean surface.

Expendable bathythermograph (XBT) records were gathered at the same time as for the acoustic data. Three XBT records taken at 21.30, 21.45, and 22.00 LT, are shown in Fig. 2. Note that the temperature profiles are essentially featureless down to the depth of the acoustic reflecting layer, at which depth small temperature gradients are visible. Previous experience[4-6] has shown that a particulate layer is often observed in coincidence with a temperature gradient. In addition to the shallowest temperature gradient (clearly visible at 45 m), the temperature profile gathered at 21.45 LT has two small temperature gradients (I and II in Fig. 2; with the corresponding reflecting layers marked I and II in Fig. 1).

From the data of Schmitz and Richardson[3], it is reasonable to expect that the northwards surface current varies from ~100 cm s^{-1} at the left of Fig. 1, to ~180 cm s^{-1} at the right of

Fig. 1 Acoustic reflecting layers and particulate columns observed while crossing the Gulf Stream: interference lines are labelled (i).



0028–0836/78/0276—0360$01.00

© Macmillan Journals Ltd 1978

Fig. 1. The mean eastwards current is very small, being maximum at the right of Fig. 1, with a magnitude no larger than ~5 cm s^{-1}.

The acoustical data shown in Fig. 1 are greatly exaggerated in the vertical; Fig. 3 is a geometrically corrected interpretation of the data between the 7.7 and 8.8 km range limits of Fig. 1. Figure 3 shows that what we are calling spires, because of their appearance of the acoustic records, are in fact more like humps or mounds. 'Hump' may not be an adequate description, however, because the along-stream extent of a particular hump may greatly exceed its cross-stream extent; in which case perhaps the term 'wall' would be more appropriate. There is acoustical evidence that the along-stream length of a hump or spire could be >1 km.

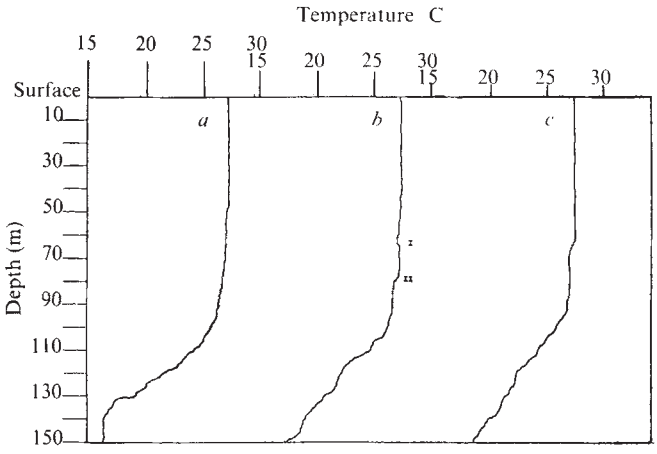


Fig. 2 Temperature profiles corresponding to the acoustic data in Fig. 1. *a*, T-11-38, 21.25LT; *b*, T-11-39, 21.43LT; *c*, T-11-40, 22.05LT.

Vertical structures

We can be certain that the vertical structures were observed in each cross-stream traverse made by the ship between longitude 28 and 25° N, but whether given spires in a crossing along-stream are continuations of given spires in another traverse is not.

The columns shown in Fig. 1 have a vertical extent, L, which is of the order of the depth of the bottom of the mixed layer, that is, if d is the depth of the bottom of the mixed layer, then

$$L \sim d \tag{1}$$

The spires have a horizontal spacing, h, which on average seems to be ~2–4 times the depth of the mixed layer:

$$2d \leqslant h \leqslant 4d \tag{2}$$

This conclusion is reached by examining substantial additional data to that shown in Fig. 1. Not all the spires are of equal size and dimensions; equation (2) is an average statement and implies that as the depth of the thermocline changes, as occurs in traversing the Gulf Stream, so too should N (the number of spires per unit horizontal distance) vary:

$$\frac{1}{4d} \geqslant N \geqslant \frac{1}{2d} \tag{3}$$

The spire width, w, at the spire base in the cross-stream direction is of the order of 2–5 times the mixed layer depth:

$$2d \leqslant w \leqslant 5d \tag{4}$$

A typical dimension for d is 50 m, so that $L = 50$ m, 100 m $\leqslant h \leqslant$ 200 m, 0.005 $m^{-1} \leqslant N \leqslant 0.01$ m^{-1}, and 100 m $\leqslant w \leqslant$ 250 m. The spire structures persist over a 24-h period and are not changed due to the rise of the deep scattering layer (DSL). In fact, the DSL rises to the thermocline depth and then contributes (for example, biota) to the particulate spires.

Table 1 Data from observations of vertical particulate spires

Date	Location	Time interval of observation	Wind speed ($m\,s^{-1}$)	Wind direction
19 Mar. 1978	14°12′ N 66°39.0′ W	20.56–22.06 LT	10	090°
11 Nov. 1977	28°59.5′ N 79°07.1′ W	08.30–13.00 LT	7–9	350–000°
11 Nov. 1977	29°03.9′ N 80°04.3′ W	19.00 LT–01.00 12 Nov.	8–10	330–350°
12 Nov. 1977	29°14.1′ N 79°42.7′ W	18.00 LT–06.30 13 Nov.	10–12	330–360°
14 Nov. 1977	26°49.3′ N 79°54.1′ W	21.00–23.30 LT	7	065–085°
31 Jan. 1973	25°21.8′ N 71°12.2′ W	11.35 LT	10	080°

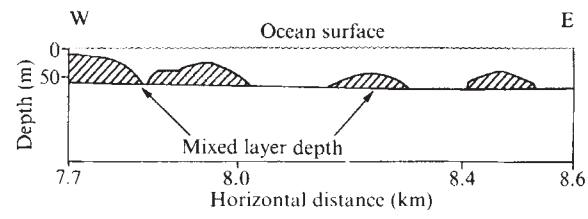
As we have stated, we are uncertain about the phenomena generating and sustaining the particulate spires. However, the hypothesis for the origin of the spires (or, perhaps walls) has two parts. First, the spires are not primarily the result of biological behaviour. The particulates comprising the reflecting layer, out of which the spires extend, are thought to be plankton, and in the case of the Florida Current, at least, are also thought to be sediment, perhaps in combination with biota. It is unlikely that such particulates are highly mobile, and, in any case, there is no report of innate particulate spire-forming behaviour for oceanic marine biota. Second, the spires arise from oceanic water circulation and congregate in spires (or walls) as a result of that circulation.

Water circulation

If the particulate structures arise from water circulation, what would be the form of that circulation? The structures might be formed by an upwards directed flow so that the particulates are carried upwards from the reflecting layer. If there is such a vertically directed flow, then only the uppermost reflecting layer contributes particulates to the flow. This can be seen by examining some of the deeper reflecting layers in Figs 1 and 4. No particulate spires extend from any other than the shallowest reflecting layer which suggests that the vertical flow is confined between the thermocline and the surface.

We feel that the proposed circulation is not due to the presence of internal waves. As may be seen from Figs 1 or 3, little or no vertical spatial variation of the shallowest reflecting layer appears on the 50–500 m length scale. Internal waves of this horizontal scale, of amplitude as small as a couple of metres, are detectable with this acoustic system[2]. However, small scale perturbations are visible on some of the deeper reflecting layers in Fig. 4. The kind of wind conditions at the time the spire observations were made may be relevant to whether circulation cells of some type are present. For example, Faller[7] has shown that certain wind and wave conditions must be present for Langmuir type circulations to occur. Table 1 shows wind and other data for various spire observations: such as measured winds exceed 5 m s^{-1} in all observations. This contradicts the

Fig. 3 Interpretation of the acoustic data in Fig. 1 from 7.7 to 8 km with the vertical exaggeration removed.



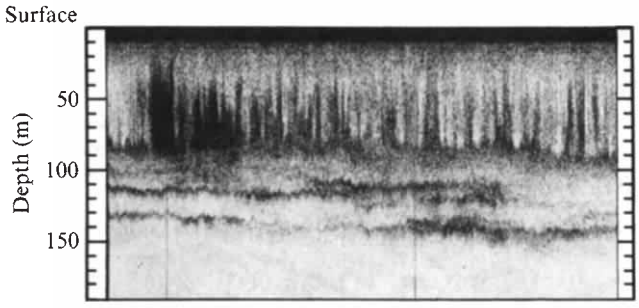


Fig. 4 Acoustic reflecting layers and spires in the area of the Antilles Current.

results reported in ref. 8: in which it is stated that the "columnar scattering" is observed in the wind mixed layer during very calm weather conditions. Table 1 also indicates that observations of spires were made during daylight and night-time conditions. We do not wish to speculate any further, questions as to whether the flows are some sort of circulation cells (so-called Langmuir cells) must be put off for the moment.

Implications of particulate spires in the Gulf Stream

One of the most important implications of the existence of particulate spires in the Gulf Stream is the small scale variability in the horizontal of particulate densities. Figure 1 shows the small scale variability of particulate densities, both in the vertical and the horizontal. While on-station, the particulate columns have the appearance of patches as the ship slowly drifts. Depending on the along-stream length of the cells, there may be a distinct difference in a biological net tow, depending on whether the tow is along-stream or cross-stream. The behaviour in the deep scattering layer, mentioned above, is interesting in that the DSL is generally thought to be mobile, but, nevertheless, preferentially aligned above the thermocline into the spires.

There are also some interesting potential implications for ocean-mixed layer studies; particularly if spire formation indicates an underlying circulation. As Pollard[9] pointed out, circulations such as those of Langmuir are potentially efficient in mixed layer momentum redistribution. A major question remains: how widespread or typical are the phenomena described here? We do not know, for example, if such spires are in some way related to the presence of ocean currents. The data gathered in Fig. 1 were clearly in a current system—the Gulf Stream. The observations shown in Fig. 4 may have been of a less well known current system[10].

Conclusion

Spires of particulate matter have been observed extending from the bottom of the mixed layer to near the surface of the ocean and in the region of different ocean current systems. The detection of the spires was made possible through the use of an ultrasensitive 20 kHz acoustical system. Further research is required into the nature of the particulates present and of the phenomena generating the spires.

Received 2 May; accepted 2 October 1978.

1. Proni, J. R. *et al. Deep Sea Res.* **23**, 8, 831–837 (1976).
2. Proni, J. R. & Apel, J. R. *J. geophys. Res.* **80**, 9, 1147–1151 (1975).
3. Schmitz, W. J. & Richardson, W. S. *Inst. Mar. Sci. Lab.* 222 (University of Miami, ML 66202 (Mimeographed) 1966).
4. Newman, F. C., Proni, J. R., Walter, D. J. & Byrne, H. M. *Nature* **269**, 790–791 (1977).
5. Drake, D. E. *Deep Sea Res.* **18**, 763–769 (1971).
6. Heathershaw, D. C. & Simpson, J. H. *Estuar. Coastal Mar. Sci.* **2**, 91–103 (1974).
7. Faller, A. J. *Science* **201**, 618–629 (1978).
8. Angel, M. V. in *Oceanic Sound Scattering Prediction* (eds Andersen N. R. & Zahuranec, B. J.) 217–248 (Plenum, New York 1977).
9. Pollard, R. T. in *Modelling and Prediction of the Upper Layers of the Ocean* (ed. Kraus, F. B.) 102–117 (Pergamon, Oxford, 1977).
10. Hansen, D. V. *A.I.A.A. Conf. Free Drifting Buoys* NASA CP-2003, Hampton, Virginia, 175–192 (1974).

Protein disk of tobacco mosaic virus at 2.8 Å resolution showing the interactions within and between subunits

A. C. Bloomer, J. N. Champness*, G. Bricogne, R. Staden & A. Klug
MRC Laboratory of Molecular Biology, Hills Road, Cambridge, UK

Protein subunits in the two layers of the disk of tobacco mosaic virus have very similar conformations. Much of the bonding between subunits is polar, including salt-bridge systems. Arginine residues play a prominent part here and elsewhere. Interactions within each layer involve groups whose contacts can be adjusted to allow the transition from disk to virus helix. Aromatic clusters within each molecule are linked in a hydrophobic girdle encircling each ring.

THE disk of tobacco mosaic virus (TMV) protein (see review by Caspar[1]), is an obligatory intermediate in the assembly of the virus from its constituent RNA and protein[2]. It fulfils both the physical requirement for nucleating particle growth and the biological requirement for specific recognition of the viral RNA. The disk is composed of two rings or layers, each containing 17 subunits[3] and both facing the same way[4]. The two layers are in contact only at high radius whereas at the opposite end of a subunit, towards the central hole of the disk, they are separated as if poised for reaction with, and intercalation of, the viral RNA[5]. Assembly is initiated by insertion into the middle of the disk[6] of a special sequence within the RNA[7] which folds into a hairpin loop[8]. Interaction with this loop has several consequences for the protein subunits: the circular subunit arrangement of the disk becomes the helical one of the virus; the two rings close about the bound RNA, losing the pairing interaction between them at high radius; the flexible part of the protein at low radius assumes the ordered structure of the virus[9]. During these rearrangements all the lateral contacts between subunits are essentially preserved while the axial ones are completely changed.

We have previously described[5], at 5 Å resolution, the subunit packing within the disk and the folding of the polypeptide chain. Extension of the analysis to 2.8 Å resolution provides details, as described here, of the conformation of the subunits and the variety of their interactions. Subunits in the two layers are remarkably similar despite their crystallographic independence. Notable features of the structure include the important structural roles of several arginine residues, the extended salt-bridge

* Present address: Wellcome Research Laboratories, Langley Court, Beckenham, Kent, UK.

0028–0836/78/0276—0362$01.00 © Macmillan Journals Ltd 1978

Table 1 Statistics of data collection

	Native	Mercury
Total reflections measured	817,000	626,000
No. of independent reflections	200,000	163,000
$R_{sym} = \Sigma_i \lvert \bar{F} - F_i \rvert / \Sigma_i F$	0.10	0.11
R_{sym}, 50% strongest reflections	0.08	0.09
Mean isomorphous difference $\langle \lvert F_{nat} - F_{der} \rvert \rangle / \langle F_{der} \rangle$	—	0.30

Using the methods of oscillation photography[13], with a range of 0.5–0.75° for a single picture, a total of about 230 film packs were used to cover the complete 90° oscillation range about the *b* axis, and the 32° of cusp data about the *a* axis, which together constitute a complete set of data. A typical film pack required an X-ray exposure of about 10 h. The crystal lifetimes varied widely up to a maximum of about 50 h exposure for any one location, with translation of the crystal giving up to three locations per crystal. Overlapping rotation ranges were used whenever the crystal location was changed. The fully and partially recorded intensities were measured with a flat-bed hybrid densitometer[14]. Films were scaled together by means of symmetry-related reflections recorded on different films and a symmetry *R* factor was computed for each compiled set of data. Reflections split between two contiguous rotation ranges were not used for scaling but are included in the statistics. They comprise 40% of the total measured reflections and are not quite as accurate as comparable fully recorded reflections, but the compiled sets of data would have been far less complete without them.

networks both within and between the rings, and the encircling of each ring by a complete girdle of hydrophobic residues.

Structure analysis

Preparation and crystallisation of the disk protein have been described previously[10]. The crystals are orthorhombic, space group $P22_12_1$, with unit cell dimensions $a = 228$ Å, $b = 224$ Å, $c = 174$ Å. There is one disk of molecular weight 600,000 per asymmetric unit[11] with its 17-fold axis approximately parallel to *c* (ref. 12).

X-ray intensities were recorded by an extension of the methods of oscillation photography[13] used for the 5 Å data and described elsewhere[5]. Almost 2×10^6 intensities were measured[14] for crystals of the native protein and a single covalently bound mercury derivative[10] (Table 1). The derivative crystals did not diffract as well at higher resolution as those of the native protein, and thus the mercury data effectively extend only to the slightly lower resolution of 2.95 Å. This was compensated for by use of the 17-fold non-crystallographic symmetry[11,12] to extend, as well as improve, the phases from the single isomorphous derivative. Heavy-atom parameters were refined independently for the 34 mercury atoms in a manner similar to that used at 5 Å resolution[5]. Details of these calculations will be published elsewhere. Phase refinement by non-crystallographic symmetry averaging[15–17] proceeded in two stages (Table 2) to investigate the existence of any departures from the 17-fold symmetry. However, no such departures were found. The averaged map resulting from stage I of this phase refinement (cycle 2 map, 1976) was used in an optical comparator[18] to build an atomic model of the protein subunit which was verified by the cycle 5 map (1977) from stage II of the refinement. This map was used both for making minor adjustments to the physical model and for model-building with a computer-graphics program (BILDER)[19]. All the atomic coordinates used in the figures accompanying this paper are taken from this computer model. Parts of the model are shown in Fig. 1 superimposed on the cycle 5 electron-density map.

Description of the subunit

The molecular boundary, the course of the polypeptide chain and the strong similarity between subunits in the two rings, which were clear at 5 Å resolution, are all confirmed at higher resolution. The chain tracing shown in Fig. 2 differs from that based on the 5 Å map only in that the wrong alternative was then chosen for the connectivity of the four-way contact at 70 Å radial distance from the centre of the disk. In no way does this affect the sequence of the four main helices, the only consequence being to reverse the direction of the chain through the 'hanging loop' from that proposed earlier.

Detailed interpretation of the map has led to an atomic model of a subunit in the B ring of the disk and all its contacts with neighbouring molecules. The densities for a subunit in the two rings of the disk are so similar that any differences in molecular conformation between the rings will be assessable only after refinement of the molecular coordinates, which is now in progress, has been completed. Examination with a computer-graphics program[19] of the model for a subunit in ring B superimposed, by a rigid-body transformation, on the density for a ring A subunit showed that the atoms in the main chain were all within 0.5 Å of the positions indicated by the density and that most of the side chains were equally close.

Table 2 Progress of the phase determination

	Stage I			Stage II			
Cycle	0	1	2	2*	3	4	5
*R factor after averaging		0.47	0.29	0.41	0.37	0.31	0.28
Correlation (between F_{obs} and F_{calc})		0.63	0.85	0.69	0.75	0.82	0.86
Figure of merit	0.48	0.63	0.78	—	—	—	—
Mean phase change from last cycle		41°	11°	41°	37°	13°	7°

*$R = \langle \lvert F_{obs} - F_{calc} \rvert \rangle / \langle F_{obs} \rangle$. Where F_{calc} is the structure factor computed from the averaged map.

The initial data set consisted of 145,000 amplitudes phased by single isomorphous replacement (SIR) with a mean figure of merit of 0.48, and of 55,000 unphased amplitudes, mostly between 3.2 and 2.8 Å resolution. Heavy-atom parameters had been refined independently for the 34 mercury sites. The parameters of the 17-fold non-crystallographic symmetry axis were then checked but found to be unchanged from those determined at 5 Å resolution. The phase refinement was done in two stages, to settle any doubt as to whether the absence of density at low radius in the 5 Å map was due to disorder or to an abuse of 17-fold averaging. For stage I the electron density was averaged only in that region of the disk lying between 40 and 90 Å radius, and was left unchanged in the rest of the asymmetric unit. This considerably weakened the phase constraints imposed by the symmetry, as the U/V ratio[16] was 0.6, but combination of phase information from SIR and from this partial averaging still acted as a constraint, and produced a marked improvement in the overall quality of the map. However, no significant departures from 17-fold symmetry could be detected, either at low radius or at contacts between adjacent disks. These unconstrained regions were rather noisy, and this noise tended to correlate with itself from one cycle to the next, leading to spuriously low residual *R* factors. For stage II, the density was 17-fold averaged in the whole disk at all radii above 20 Å, and was reset to its average value outside the disk boundaries. Deep holes of negative density at the heavy-atom sites, due to poor SIR parameters, were removed from the maps. The phases calculated from such maps were associated directly with the observed amplitudes, without weighting and without recombination with the isomorphous phases, to produce the structure factors used in the next cycle. The use of this new procedure, starting from the phases of cycle 2, induced a large readjustment of the poorer phases (cycle 2*), followed by convergence. The constraints imposed by the 17-fold symmetry were 13.5 times as strong ($U/V = 0.045$) as in stage I, yet a comparably low residual factor was obtained. This strongly suggests that all the ordered density in the disk accurately obeys the 17-fold symmetry.

The central part of the molecule consists of a group of four α helices which are arranged closely parallel or anti-parallel to each other and extend between 45 and 65 Å radius[5]. The left and right slewed helices, LS and RS, in the upper half of the subunit are joined at the end proximal to the disk axis by a short bend in the protein backbone referred to as the slewed hairpin. The right and left radial helices, RR and LR, in the lower half of the subunit both fade into weak discontinuous density, indicative of the flexibility of this loop region of the protein in the

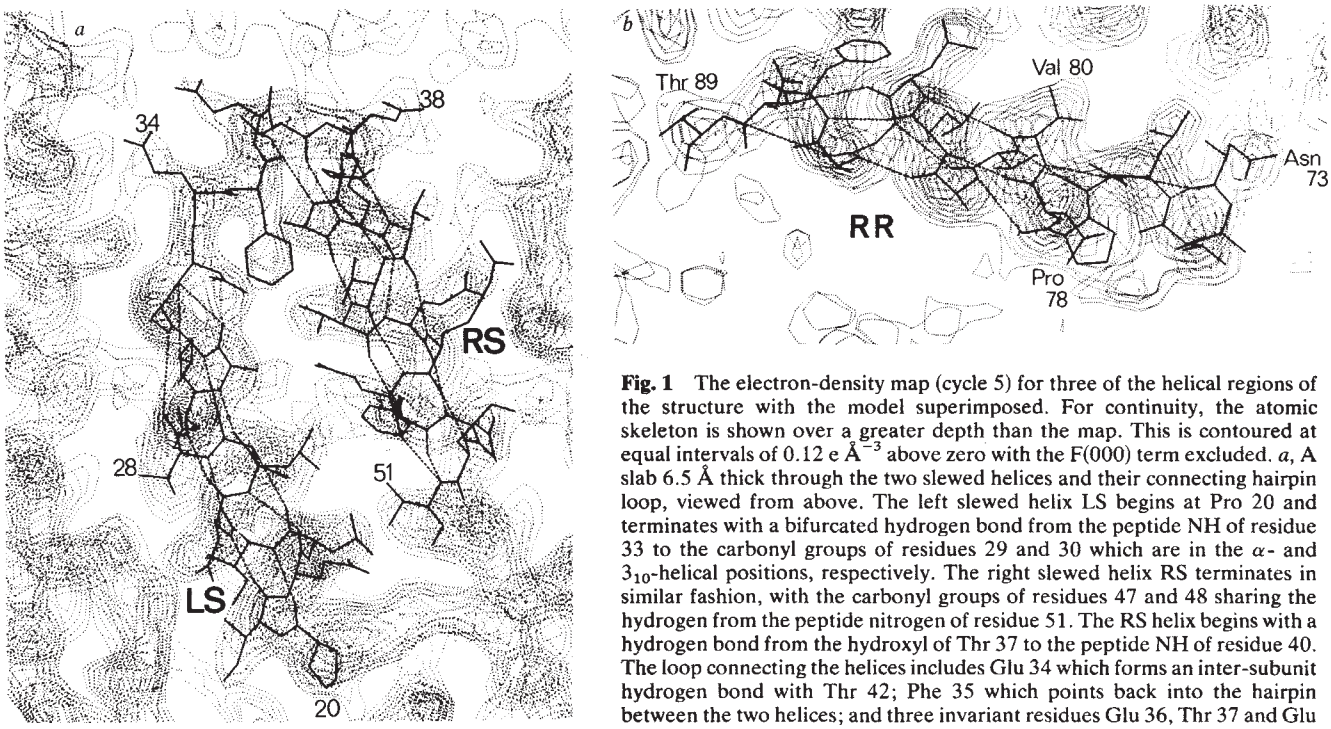


Fig. 1 The electron-density map (cycle 5) for three of the helical regions of the structure with the model superimposed. For continuity, the atomic skeleton is shown over a greater depth than the map. This is contoured at equal intervals of 0.12 e $Å^{-3}$ above zero with the F(000) term excluded. *a*, A slab 6.5 Å thick through the two slewed helices and their connecting hairpin loop, viewed from above. The left slewed helix LS begins at Pro 20 and terminates with a bifurcated hydrogen bond from the peptide NH of residue 33 to the carbonyl groups of residues 29 and 30 which are in the α- and 3_{10}-helical positions, respectively. The right slewed helix RS terminates in similar fashion, with the carbonyl groups of residues 47 and 48 sharing the hydrogen from the peptide nitrogen of residue 51. The RS helix begins with a hydrogen bond from the hydroxyl of Thr 37 to the peptide NH of residue 40. The loop connecting the helices includes Glu 34 which forms an inter-subunit hydrogen bond with Thr 42; Phe 35 which points back into the hairpin between the two helices; and three invariant residues Glu 36, Thr 37 and Glu 38 which all point inwards towards low radius where they are exposed to the solvent. *b*, A slab 5 Å thick through the right radial helix RR, viewed from the side, showing Pro 78 contained within this helix which starts at residue 74 with a bifurcated hydrogen bond from the peptide carbonyl of 73. The proline prevents the formation of two potential hydrogen bonds within the helix, those which would have involved the NH groups of residues 78 and 79, and causes the helix axis to become more closely horizontal at higher radius. The two peptides whose carbonyl groups are unable to form hydrogen bonds are tilted away from the regular helical conformation, exposing them to the solvent. The helix continues to Asp 88 where the density then becomes very faint, with Thr 89 probably not in a helical conformation but only weakly indicated in this map.

absence of RNA[5]. The flexible loop which connects these radial helices at their proximal ends is, in the virus, located at lower radius than any other part of the protein[9]. The distal ends of all four helices are braced together by a short anti-parallel β sheet of four strands. Amino acid side chains on the high-radius side of this sheet form part of a hydrophobic girdle which extends continuously around the circumference from one subunit to the next. The girdle also includes residues from the upper loop, a piece of extended chain between residues 54 and 67, joining the RS and RR helices, the hanging loop extending from the N terminus towards residue 16 of the β sheet and the RS helix, and the C-terminus region arising from the β bend at the end of the LR helix. Within this girdle is an extended aromatic cluster in the centre of each subunit and a smaller such cluster across the subunit interface. Elsewhere in this interface and that between the two layers, there are systems of salt bridges involving some of the many arginine residues present in TMV protein. Arginine residues play an important part in both intra- and inter-subunit interactions, exploiting fully the high specificity and high directionality of the bonds which can be formed. These various elements of the structure will be considered in turn.

The α-helical core and the β sheet

The four main α helices include about 60 residues out of the total of 158 (ref. 20), with residues 20–32, 38–48, 74–88 and 114–134 occurring in the LS, RS, RR and LR helices, respectively. Residues 49–52 form a 3_{10}-helical extension of RS at its distal end. All the helices are somewhat distorted from the idealised structures. The helices are packed as in a segment of a left-handed, 4-stranded supercoiled bundle with hydrophobic contacts along the centre. This hydrophobic core contrasts with the many polar and charged groups on the outside of the supercoil where there are alternating patches of polar and hydrophobic residues. The ends of the helices are orthodox, with bifurcated hydrogen bonds finishing off both LS and RS (Fig. 1*a*) and starting off RR (Fig. 1*b*), a hydrogen bond from a side chain at the beginning of RS and a β bend at the end of LR. Whereas Pro 20 starts the LS helix in the usual manner, the position of Pro 78 within the RR helix is highly unusual. The proline distorts the helix, causing its axis to incline more steeply upwards towards lower radius, thereby preventing the formation of two of the helical hydrogen bonds (Fig. 1*b*).

The hairpin bend joining the slewed helices contains the highly conserved group of residues 33–38. Phe 35 extends back inside the hairpin with its aromatic ring between the pair of helices in a similar manner to the packing observed in myoglobin of Phe 123 between the G and H helices[21]. Invariant glutamines 36 and 38 extend inwards towards low radius, suggesting an availability for interaction with nucleic acid. Other residues in this loop form bonds to the lower half of the subunit. The flexible loop of residues 89–113 connecting the radial helices is ordered only in the presence of RNA and so the map has no interpretable density at 20–40 Å radius, where these residues would be expected, and only weak density for residues 114–116 in the LR helix. When the mobility of the loop[22] is reduced by nucleotide binding to the disk, it seems, as discussed below, that the LR helix begins properly at residue 109 or 110. The RR helix terminates abruptly after Asp 88, which forms an inter-subunit salt bridge with Arg 122.

The distal ends of the four helices are connected transversely by a narrow and twisted strip of β sheet shown schematically in Fig. 3. The four main strands of the sheet tie the helices together in pairwise fashion with the two slewed helices above linked to the upper two strands of the sheet and the two radial helices below to the lower strands. The two central strands of the β strip are themselves connected through the high-radius upper loop. This loop and the other connecting loops in the sequence do have some very short helical segments (α and 3_{10}) but few of

these are long enough to have even a single helical hydrogen bond.

The hydrophobic girdle

The central part of each subunit on the distal side of the β sheet is a cluster of aromatic residues: Trp 17, Tyr 70 and Tyr 139 from the sheet, Phe 12 and Phe 144 from the two terminal loops, and Phe 62 from the upper loop. These groups, together with Pro 56 and Pro 63, form an extended cluster in which the rings are mostly in an edge-to-face arrangement. At the same radius of 75 Å, but extending across the molecular interface, there is a further cluster composed of Pro 20 and Phe 67 with Pro 7 and Phe 10 of the neighbouring molecule. These two clusters are joined by several hydrophobic aliphatic side chains, arising from elements of all four chains which cross this radius, to give a continuous belt of hydrophobic interactions encircling each ring of the disk.

Salt-bridge systems

The subunits in the two layers have three regions of interaction (Fig. 4) of which the largest is the extended salt-bridge system shown in Fig. 5. This involves the only two lysine residues and four acidic groups, from three segments of the main chain in the lower ring with Asn 127, Glu 131 and Arg 134 from the LR helix in the upper ring. Almost all the possible hydrogen-bonding capabilities are utilised in this complex 3-dimensional network, in which two water molecules also participate. Note that the disk is stable in very high ionic strengths[23,24], so that a simple view of the instability, in high ionic strengths, of isolated salt bridges does not apply to an extended system like this with hindered access to solvent.

Adjacent subunits within a ring are linked at low radius by a simpler salt-bridge system involving Arg 122 and Asp 88. In this case also, most of the hydrogen-bonding potential of the arginine is utilised, with additional links to Thr 89 in the neighbouring subunit and to the peptide carbonyl groups of residues 31 and 33 in the same subunit.

At higher radius in the same molecular interface (Fig. 2), the region of polar interaction has an intra-subunit salt link involving Arg 71 and Asp 77 with additional hydrogen bonds to Thr 81, Ser 49, and an inter-subunit link to Thr 28. In the cowpea strain of the virus, however, coordinated sequence changes[25] result in this region also having an inter-subunit salt linkage.

Role of arginine residues

The basic residues in TMV comprise 2 lysine, 11 arginine and no histidine residues[20]. The role of the lysine residues and three arginine residues (134, 122 and 71) in various salt-bridge systems of interaction within the subunit interfaces has been discussed above. In the subunit itself there are also two simple salt linkages involving Arg 141 with Asp 64 and the invariant pair Arg 61 with Glu 145. Both these pairs are exposed on the outer circumference of the disk where they link the C-terminal and upper loops of the protein.

Interaction with RNA is expected to involve three arginine residues[1] out of the four invariants (41, 90, 92, 113) which occur at low radius. Arg 41, the only such residue seen in our map, seems less likely to have such an interaction than the other three whose positions, as extrapolated by model-building from positions 89 and 114, would have them all pointing down from the A ring into the 'jaws' between the layers of the disk.

The high specificity and directionality of the polar bonds formed by arginine residues contrast with the hydrophobic surfaces presented by their methylene chains. That of Arg 61 lies against the planar face of Trp 152, and similarly Arg 122 is close to Leu 31, and Arg 41 lies between the rings of Phe 35 and Phe 87.

Inter-subunit contacts

Contacts between the layers: Three regions of contact are indicated in Fig. 4, all of different type and all occurring at the higher-radius end of the molecule but spaced apart. The outermost is a polar interaction between serines 147 and 148 and Thr 59. At lower radius are the extended salt-bridge system described above and also a small hydrophobic patch where Ala 74 and Val 75 both touch the ring of Pro 54. These three contacts form an essentially kinematic mounting of a subunit of

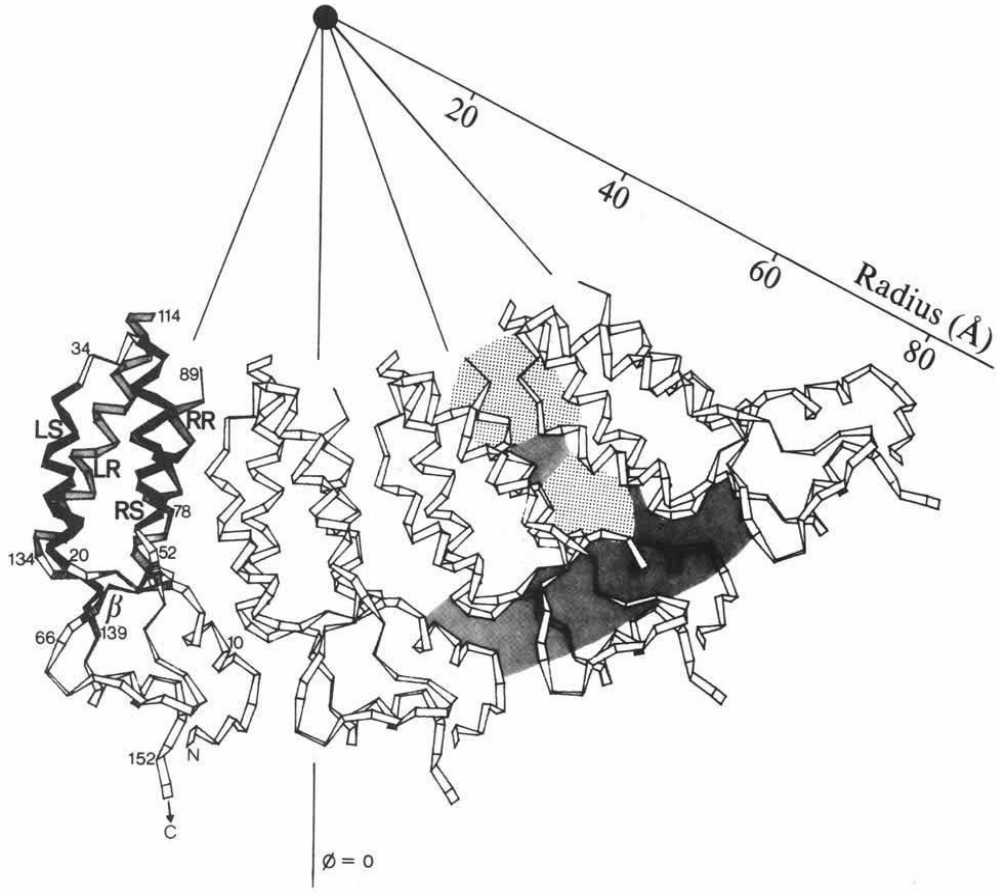


Fig. 2 The folding of the polypeptide chain and the lateral interactions between subunits. Four adjacent subunits within a ring are shown viewed from above. The N and C termini of the chain are marked on the extreme left subunit together with the four main helices, left and right slewed (LS and RS) in the upper half of the molecule with right and left radial (RR and LR) below, and the β sheet which connects all four helices at their higher radius ends. The other subunits show the overall shape of the subunit as a scimitar, with a pronounced slew relative to the radius of the ring. The interface contains alternating patches of polar residues, indicated by stippling, and of hydrophobic residues, indicated by solid shading, which are all contributed to by residues from both the subunits forming the interface. The hydrophobic patch at higher radius is continuous with the hydrophobic girdle which extends circumferentially across the whole width of a subunit and across the interface, forming a continuous belt around the ring. The hydrophobic girdle includes residues from the high radius side of the β sheet, the upper loop which joins the RS and RR helices and the two lower loops leading to the termini of the chain.

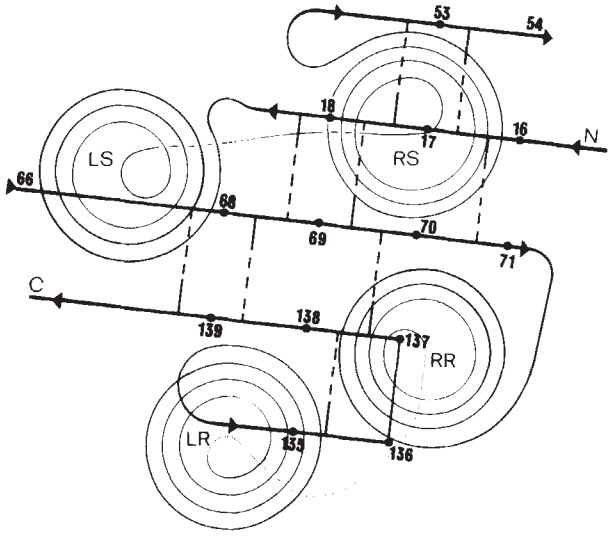


Fig. 3 Schematic view of the anti-parallel β sheet showing the connectivity of the strands and of the four principal α helices. The direction of view is along the long axis of the subunit towards the centre of the disk. The β sheet is a narrow, twisted strip which extends across the subunit bracing together the four α helices. At the bottom of the β sheet is a β bend which occurs immediately after the end of the LR helix. The bend is of type II where the second residue is commonly glycine and indeed Gly 137 is invariant in all known strains of TMV. The residues whose side chains are on the higher-radius side of the sheet are the two lysines (53 and 68), which form part of an extended salt bridge (Fig. 5), and three aromatic residues (Trp 17, Tyr 70 and Tyr 139) which form part of a cluster within the continuous hydrophobic girdle.

one ring onto one of the other. They are completely changed during the transition from disk to helix.

Contacts within a layer: By contrast, the four patches of alternating polar and hydrophobic contacts (Fig. 2) are maintained during the disk to helix rearrangement. There is no evidence of any relative motion within a subunit during this transition, and the coordinates of heavy-atom labels at four equivalent sites[26] in the disk and the virus show that the molecule moves as a rigid body[24]. Considerable flexibility is thus required in the side-chain contacts if the backbone conformation stays essentially rigid. During the transition, subunits in the A and B rings tilt[27] by 10° and 20°, respectively, about a tangential axis[5,9], so changing by 4° and 8° the hinge angle at the interface between neighbouring subunits, if these are envisaged as being related by a hinge extending radially along the interface. This change in hinge angle could mean an increased distance between the slewed helices in the upper part of the interface, a decreased distance between the radial helices in the lower part, or both of these. The first alternative is the least likely, as the transition results in a more compact structure[6] in which the subunits are closer to the central axis and the number of subunits per turn is decreased from 17 in the disk to $16\frac{1}{3}$ in the virus. Figure 2 shows that in the disk the interface has more space between the radial than the slewed helices to allow the transition. Both the polar patches in the interface have residues which act as spacers: Arg 122 in the patch proximal to the axis and Tyr 72 in the distal patch. Rearrangement of the packing of these side chains allows the necessary movement in the polar patches. The non-directional character of the interactions in the hydrophobic patches could allow the apposed surfaces there to slide over each other during the transition.

Comparison with virus

In describing the structure of the RNA and its binding site, as deduced from a map of the virus itself using 4 Å data, Stubbs *et al.*[9] presented a model for the four main α helices which differs in several respects from that presented here. The largest differences are in the radial helices. Tyr 139 is located by a label in each model[24] and, whereas our model has a compact β bend separating this from Arg 134, the virus model has a greater separation. Comparison of our map with the 5 Å resolution map of the disk containing bound oligonucleotides[28], and computer model-building[19] indicate that, in the presence of bound nucleotide, the LR helix continues inwards to residue 109 or 110 before the chain turns upwards into the vertical column of density (V) seen in the virus map[9]. In the virus model this turn occurs at position 113, a difference of +3 or +4 residues between our model and the virus model. In our RR helix, Pro 78 occurs right at the bottom of the subunit marking a turn where the helix axis becomes more closely horizontal with increasing radius, whereas in the virus model the chain becomes less horizontal and then continues out in radius to residue 70. In our model we place Tyr 72 at the end of the radial density and residues 71 and 70 lie much higher than this radial helix (Fig. 4). None of the helices in the disk model have the idealised geometry which was built into the virus model, with the result that, although the two

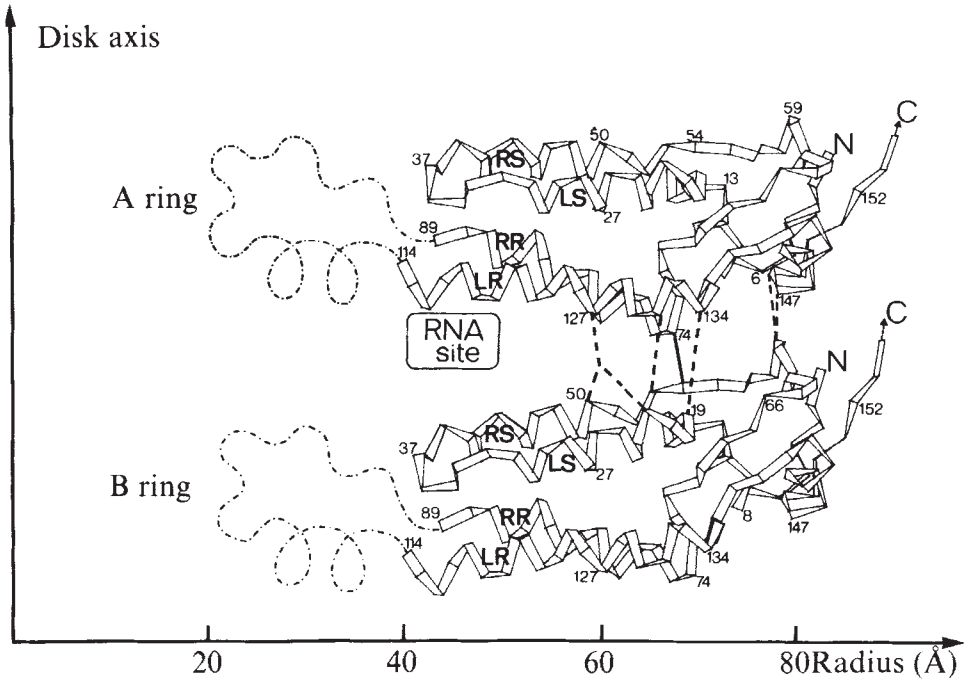


Fig. 4 Side view of a sector through the disk showing the relative disposition of subunits in the two rings and the axial contacts between them. There are three regions of contact indicated by a solid line for the hydrophobic contact of Pro 54 with Ala 74 and Val 75, by dashed lines for the hydrogen bonds between Thr 59 and serines 147 and 148, and further dashed lines for the extended salt-bridge system shown in greater detail in Fig. 5. The low-radius region of the chain, which has no ordered structure in the absence of RNA, is shown schematically with an indication of an additional 1–2 turns extending the LR helix (see text) before it turns upwards into the vertical column. The primary nucleotide-binding site is probably below the RR helix[28].

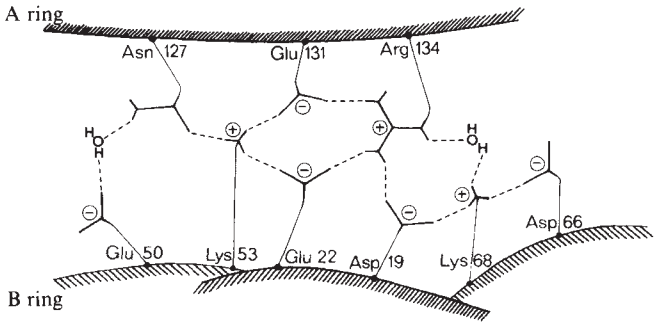


Fig. 5 The extended salt-bridge system between subunits in the two rings of the disk presented schematically in a side view. All three residues in the A ring emanate from the LR helix whereas those in the B ring arise from the 3_{10} extension of the RS helix and the top strand of the β sheet (50 and 53), the beginning of the LS helix (19 and 22) and lastly, the end of the upper loop and the third strand of the β sheet (66 and 68). The salt-bridge system, which utilises almost all possible hydrogen-bonding sites, is a complex three-dimensional network which is only presented here schematically. There are clear features in the electron-density map to represent both water molecules.

polypeptide chains may be out of phase by −2 residues beyond the distal end of the RR helices, both models can still have, for example, Asp 88 pointing out across the subunit interface. In the LS helix Cys 27 is located by a mercury site in both the disk[12] and the virus[29] but, whereas in our model Cys 27 points downwards towards the centre of the four-stranded supercoil, in the virus model it faces the RS helix[9], indicating a difference of +1 residue between the models.

The RR helix in the virus model continues for two or three turns beyond Asp 88 where, in the disk, it terminates at the start of the flexible loop. Extension of the RR helix beyond this point is not possible in our model because of interference from residues 38 and 41. Preliminary evidence from a low resolution study of oligonucleotides bound to the disk[28], while suggesting that the LR, LS and RR helices all move slightly upwards, still gives no clue as to how the flexible loop is folded. However, the different assignments of phase in the α helices mean, first, that our model has fewer residues to be fitted in between this point and the start of the LR helix and thus they need not be as tightly packed as the α helices of the virus model; and second, that the existence of a carboxyl cage[9], so-called in the virus model, must be treated with caution although, purely on the basis of the sequence, there must be a concentration of negative charges in this region of the molecule.

Discussion

The protein of TMV is the only protein from a large macromolecular assembly whose structure is now known at this level of detail. The subunit is a small protein folded into a single domain within which the folding of the amino and carboxyl halves of the polypeptide chain is quite similar. Between these two parts of the molecule there is an approximate twofold symmetry, which is geometrical as well as topological, indicative of a possible gene duplication (A. D. McLachlan, personal communication). The axis of this symmetry is along the centre of the four-stranded supercoil and approximately through the centre of the β sheet. Variation in the character of the protein along the direction of this axis reflects a segregation of function. Nearest to the centre of the disk is the flexible loop, awaiting interaction with RNA. Next comes the α-helical core and β sheet providing rigidity within the subunit while allowing flexibility across the interface. Last, at the distal end of the subunit is the hydrophobic girdle extending around the entire ring.

The discovery of the flexible loop in our 5 Å map[5] gave insight into the means whereby TMV protein can interact with, in effect, a random RNA sequence in the complete virus, and yet also recognise the special nucleation sequence of TMV RNA[7]. The binding site must combine the specificity required for the initial recognition of the special sequence with the ability to accommodate subsequently any sequence of RNA. Stereospecificity will be provided by the rigid parts of the protein in the vicinity of the RNA binding site: the proximal ends of the RR and LR helices and the hairpin bend, joining the proximal ends of the two slewed helices, which forms a rigid wall (with a concentration of invariant residues) facing into the cavity of the disk. Preliminary evidence suggests that nucleotides bind below the RR helix[28].

The mobility of the flexible loop in the disk is reminiscent of that found in the activation domain in trypsinogen[30]. Segments from four separate elements of this chain are flexible in the zymogen but become tightly interdigitated in the enzyme, after cleavage of the activation peptide and associated conformational changes, forming a rigid, correctly defined specificity pocket for binding the substrate. Creation of the binding site in TMV does not involve protein cleavage, as observed in trypsin, but it may be noted that when such cleavage does occur in the flexible region of disks of TMV it is essentially nonproductive, leading irreversibly to the stacked disk aggregate[31] of TMV protein which is not capable of initiating the assembly process.

The strong tendency of TMV protein to aggregate[1] can now be understood in terms of its surface properties. The hydrophobic nature of both the centre of the supercoiled bundle formed by the four α helices and the aromatic cluster in the centre of the subunit on the opposite side of the β sheet are not unusual, but the hydrophobic patches on the molecular surface would only be expected in a component of a larger assembly, not in a monomeric protein. These patches occur only within the lateral interface, which is essentially the same in the disk and the virus. This interface comprises hydrophobic and salt interactions which would both be favoured by higher temperatures, and indeed, TMV is a classical example of an entropically driven system[32]. The general character of the alternating hydrophobic and salt-bridge patches within this interface is conserved in all strains of the virus which have been sequenced, although the boundaries of the patches differ slightly between strains. A fuller discussion of this will be presented elsewhere.

The salt-bridge systems within the lateral interface remain on the transition from disk to virus, but the most extensive system, that between the layers of the disk, is completely disrupted by the transition, which changes the relative positions of subunits in the two layers both axially and azimuthally[5]. The interaction between subunits in successive turns of the virus places the first turn of the RR helix, which is the lowest part of the subunit in the A ring of the disk (Fig. 4), in the groove between the distal ends of the LS and RS helices.

The implications for virus assembly of this rearrangement were discussed in our previous paper[5]. The prediction that RNA interacts with TMV protein from the centre of the disk has been verified with the discovery of a hairpin loop structure[8] for the nucleation sequence within the RNA[7] and the observation that both tails of the RNA emerge from the same end of the growing rodlet[33,34]. Insertion of this loop into the central hole of the disk[6], and incorporation of RNA between the two layers, resolve the paradox posed by this two-layered structure. There is now no need to postulate separate mechanisms for the initiation and elongation stages of virus assembly. A travelling loop within the RNA which starts at the nucleation sequence for the initiation process and moves towards the 5′-end during elongation, uses the same forces of interaction for both processes. An understanding of this interaction in atomic detail must await the correlation of our model for the protein with the high-resolution fibre-diffraction pattern of the virus. This study is now in progress (in collaboration with Dr G. J. Stubbs).

We thank Drs P. J. G. Butler and J. Graham for help and for discussions, Dr R. Diamond for advice and the use of his graphics program, Mrs C. Harley for preparation of the crystals, Professor K. C. Holmes and the group in Heidelberg for a copy of their virus map and for showing us their atomic model, Trinity

College for the award of a Research Fellowship to G.B. and the Cambridge University Computing Service for making special arrangements for all the necessary computations. We also acknowledge the earlier contribution of the late Dr A. Jack (1948–78).

Received 24 August; accepted 2 October 1978.

1. Caspar, D. L. D. *Adv. Protein Chem.* **18,** 37–118 (1963).
2. Butler, P. J. G. & Klug, A. *Nature new Biol.* **229,** 47–50 (1971).
3. Finch, J. T., Leberman, R., Chang, Y.-S. & Klug, A. *Nature* **212,** 349–350 (1966).
4. Finch, J. T. & Klug, A. *Phil. Trans. R. Soc.* B**261,** 211–219 (1971).
5. Champness, J. N., Bloomer, A. C., Bricogne, G., Butler, P. J. G. & Klug, A. *Nature* **259,** 20–24 (1976).
6. Butler, P. J. G. *et al. Proc. 3rd John Innes Symp.*, 101–110 (1976).
7. Zimmern, D. & Butler, P. J. G. *Cell* **11,** 455–462 (1977).
8. Zimmern, D. *Cell* **11,** 463–482 (1977).
9. Stubbs, G. J., Warren, S. G. & Holmes K. C. *Nature* **267,** 216–221 (1976).
10. Leberman, R., Finch, J. T., Gilbert, P. F. C., Witz, J. & Klug, A. *J. molec. Biol.* **86,** 179–182 (1974).
11. Finch, J. T., Gilbert, P. F. C., Klug, A. & Leberman, R. *J. molec. Biol.* **86,** 183–192 (1974).
12. Gilbert, P. F. C. & Klug, A. *J. molec. Biol.* **86,** 193–207 (1974).
13. Arndt, U. W. & Wonacott, A. J. (eds) *The Rotation Method in Crystallography* (North-Holland, Amsterdam, 1977).
14. Mallett, J. F. W., Champness, J. N., Faruqi, A. R. & Gossling, T. H. *J. Phys. E: Sci. Instrum.* **10,** 351–358 (1977).
15. Jack, A. *Acta crystallogr.* A**29,** 545–554 (1973).
16. Bricogne, G. *Acta crystallogr.* A**30,** 395–405 (1974).
17. Bricogne, G. *Acta crystallogr.* A**32,** 832–847 (1976).
18. Richards, F. M. *J. molec. Biol.* **37,** 225–230 (1968).
19. Diamond, R. *Proc. IUPAB Symp. biomolec. Struct.*, Madras (1978).
20. Dayhoff, M. O. *Atlas of Protein Sequence and Structure*, Vol. 5, D283 (National Biomedical Research Foundation, Washington, 1972).
21. Takano, T. *J. molec. Biol.* **110,** 537–568 (1977).
22. Jardetzky, O., Akasaka, K., Vogel, D., Morris, S. & Holmes, K. C. *Nature* **273,** 564–566 (1978).
23. Durham, A. C. H., Finch, J. T. & Klug, A. *Nature new Biol.* **229,** 37–42 (1971).
24. Graham, J. & Butler, P. J. G. *Eur. J. Biochem.* **83,** 523–528 (1978).
25. Rees, M. W. & Short, M. N. *Biochim. biophys. Acta* **393,** 15–23 (1975).
26. Graham, J. thesis, Univ. Cambridge (1977).
27. Wilson, T. M. A., Perham, R. N., Finch, J. T. & Butler, P. J. G. *FEBS Lett.* **64,** 285–289 (1976).
28. Graham, J. & Butler, P. J. G. *Eur. J. Biochem.* (in the press).
29. Franklin, R. E. & Holmes, K. C. *Acta crystallogr.* **11,** 213–220 (1958).
30. Huber, R. & Bode, W. *Acc. Chem. Res.* **11,** 114–122 (1978).
31. Durham, A. C. H. *FEBS Lett.* **25,** 147–152 (1972).
32. Lauffer, M. A., *Entropy Driven Processes in Biology* (Springer, Heidelberg, 1975).
33. Butler, P. J. G., Finch, J. T. & Zimmern, D. *Nature* **265,** 217–219 (1977).
34. Lebeurier, G., Nicolaieff, A. & Richards, K. E. *Proc. natn. Acad. Sci. U.S.A.* **74,** 149–153 (1977).

Tomato bushy stunt virus at 2.9 Å resolution

S. C. Harrison, A. J. Olson, C. E. Schutt* & F. K. Winkler†

Gibbs Laboratory, Harvard University, 12 Oxford Street, Cambridge, Massachusetts 02138

G. Bricogne

MRC Laboratory of Molecular Biology, Hills Road, Cambridge, UK

The polypeptide chain of a TBSV subunit folds into two domains, connected by a hinge, and a flexibly-linked N-terminal arm. Sixty of the 180 N-terminal arms interdigitate in groups of three, in an unexpected mode of protein association. The remaining 120 arms are not uniquely positioned with respect to the rest of the subunit. RNA is also not uniquely fixed to sites on the major domains.

SMALL, regular viruses self-assemble in solution[1,2]—implying that the protein subunits can associate correctly without further direction and that they can package nucleic acid. In most spherical viruses, correct assembly requires that a single type of subunit be capable of several, somewhat different bonding relations ('quasi-equivalence'). The structure of tomato bushy stunt virus (TBSV), as it appears at 5.5 Å resolution[3], shows that its subunit (MW ~ 40,000) achieves such flexibility by having two rigidly-folded domains with an interconnecting hinge. Two states of the hinge, differing by about 20°, are present in the $T = 3$ (180-subunit) icosahedral structure. A large number of inter-subunit contacts are conserved in this way, maintaining specificity.

The low-resolution view of TBSV left a number of questions unanswered. How is the hinge flexibility controlled and a particle of correct size generated? With what precision are inter-subunit contacts conserved? How is RNA packed? To explore these questions, we have extended the resolution of the TBSV structure determination to 2.9 Å. We present here a first description of the results, which show several remarkable features. Near the N terminus of the polypeptide chain, 60 of the 180 chains are ordered, interdigitating in groups of three, whereas the remaining 120 extend in a less precise way towards the particle centre. Moreover, RNA does not seem to bind to the part of the subunit rigidly fixed in the surface lattice, and the structure suggests that the N-terminal arm, flexibly-linked to the rest of the subunit, has an RNA-binding function.

Structure determination

TBSV crystallises in the cubic space group I23, $a = 383.2$ Å (refs 3–5). Diffraction to 2.85 Å was measured by oscillation photography[6], with a $\frac{1}{2}°$ range for each picture and with a total oscillation range of 25° about [110] required for a full data set. By translating a crystal between exposures, we could generally obtain two good photographs, and we therefore needed about 25 crystals for complete data. The films were scanned using a modified version of SCAN 12[7] to control an Optronics Photoscan interfaced to a PDP 11/20. As about 75% of the reflections were partially recorded, and as the crystal lifetime in any one location did not permit summing intensities of complementary spots on adjacent films in a series[6], a strategy for correcting to fully-recorded equivalents was adopted: details of the method will be published elsewhere (F.K.W. and C.E.S., in preparation). Corrected intensities were included in the final data set from only those reflections on any film that were at least half-recorded: by taking photographs with approximately abutting oscillation ranges, essentially all reflections were measured on one film or the other. The accuracy of measurements corrected in this way is comparable to the accuracy of fully-recorded reflections of the same observed strength.

Three data sets were collected: native and two derivatives (UO_2^{2+}, Hg^{2+}). The complete native data set contained 198,000 reflections: some statistics are given in Table 1*a*. Results of the heavy-atom refinement are shown in Table 1*b*. Phase determination by non-crystallographic symmetry averaging[8] proceeded in a manner similar to the phase refinement at 5.5 Å (ref. 3). A full account of the calculations, carried out at the CECAM Workshop on Virus Crystallography, will appear elsewhere (G.B., in preparation). A brief summary is included in Table 1*b*. The Fourier map used for averaging[8] was computed on a 0.6 Å grid from 182,000 terms between 16 and 2.9 Å; it had 37×10^6 points in an I222 asymmetric unit. The map computed after four cycles of phase refinement was used for the computation of

* Present address: MRC Laboratory of Molecular Biology, Hills Road, Cambridge, UK.
† Present address: EMBL, Heidelberg, FRG.

0028–0836/78/0276—0368$01.00

© Macmillan Journals Ltd 1978

Table 1 Details of the structure determination of TBSV

a. Data collection:

Form	Reflections measured	R
	6 Å → 2.89 Å	
Native	182,000	0.13
	(plus 17,000 to 6 Å from earlier work)	
UO_2^{2+}	164,000	0.14
Hg^{2+}	130,000	0.15

$$R = \frac{\sum_h \sum_{j=1}^{N_h} |I_{hj} - \bar{I}_h|}{\sum_h N_j \bar{I}_h},$$

where the reflection, h, has been measured N_h times in data collection, and the average and differences are taken after appropriate scaling; and

$$\bar{I}_h = \sum_{j=1}^{N_h} W_{hj} I_{hj} \Big/ \sum_{j=1}^{N_h} W_{hj}$$

where $W_{hj} = 1/\sigma^2(I_{hj})$. The mean value of N_h for the native data set was 3.

b. Phase determination:

Heavy-atom parameters (re-refined with phases after one cycle of noncrystallographic-symmetry phase refinement, using selected well-measured reflections; see ref. 3).

UO_2^{2+}						Hg^{2+}					
Site	F	B	x	y	z	Site	F	B	x	y	z
U A1	83	22	−0.0446	0.0708	0.3363	H A1	78	20	−0.0377	0.0896	0.3336
U B1	74	21	0.0225	0.1082	0.3383	H B1	91	16	0.0357	0.0931	0.3340
U C1	81	21	0.0225	0.0304	0.3364	H C1	74	21	0.0019	0.0271	0.3350
U A2	91	19	−0.0147	0.0979	0.3477	H A2	37	26	−0.0677	0.1208	0.4233
U B2	84	18	0.0316	0.0676	0.3471	H B2	39	27	0.0815	0.0977	0.4226
U C2	43	27	−0.0171	0.0437	0.3458	H C2	37	26	−0.0150	0.0151	0.4307
U C3	44	15	0.0172	0.0261	0.3500						
			$\langle F_{nat}\rangle = 84.3$						$\langle F_{nat}\rangle = 84.3$		
			$\langle \Delta F\rangle = 16.3$						$\langle \Delta F\rangle = 13.9$		
			$R_c = 0.655$						$R_c = 0.662$		
			$\langle F_c\rangle/E_{r.m.s.} = 1.71$						$\langle F_c\rangle/E_{r.m.s.} = 1.38$		

R_c = Cullis R factor for centric reflections

$\langle F_c\rangle/E_{r.m.s.}$ = 'phasing power' (mean calculated F divided by r.m.s. lack of closure)

Noncrystallographic-symmetry phase refinement[8]

Cycle	R	$\langle\lvert\Delta\phi\rvert\rangle$	m_{MIR}	m_{Sim}	m_{comb}
1	0.41	53°, 33°*	0.46	0.56	0.70
2	0.30	25°			
3	0.23	8.5°			
4	0.20	5.1°			

* Cycle 1 was performed with phase combination; the mean $\Delta\phi$ between MIR and calculated phases was 53°, the mean $\Delta\phi$ between MIR and combined phases was 33°. Subsequent cycles were carried out with direct phase transfer.

R: $\frac{\sum \lvert\, |F_{calc}| - |F_{obs}| \,\rvert}{\sum |F_{obs}|}$, where F_{calc} is the amplitude calculated from a symmetrised map at each stage, and F_{obs} is the measured amplitude

$\langle\lvert\Delta\phi\rvert\rangle$: mean phase change

m_{MIR}, m_{Sim}, m_{comb}: mean figures-of-merit for cycle 1

icosahedrally averaged maps of individual subunits, using a generalised skew planes procedure[8].

In considering the effects of icosahedral averaging, both in phase refinement and in examining the final maps, it is important to recall that the TBSV particles lie at positions of tetrahedral symmetry in the crystals. There is no chemical evidence for 'twelveness' in the particle—that is, no evidence for any special role for the tetrahedral subgroup of symmetry operations[9]. Therefore, the fivefold averaging imposed by the calculation is not 'smearing out' any structures (such as RNA or the possible unique subunit) that have not already been averaged out by the space group itself. Tetrahedral distortions of the particle, imposed by the lattice contacts, are of course always possible. But a difference map using $(|F_{obs}| - |F_{calc}|) \times \exp(i\alpha_{calc})$ coefficients, 'calc' referring to the last cycle of icosahedral averaging, showed no significant features, even near interparticle contacts. Such distortions must therefore be extremely small.

Maps of subunits in quasi-equivalent locations A, B and C (Fig. 1) were prepared, sectioned such that all three domains appear in the same orientation, using superposition parameters determined from the 5.5 Å map[3]. This presentation facilitates visual comparisons. Features in these averaged maps are extremely clean and well defined (Fig. 2): the polypeptide chain can be traced unambiguously, side chains are well-outlined, and main-chain carbonyls are visible for most residues. The structural results to be presented below are all derived from inspection and measurement of mini-maps (3.2 mm per Å) contoured from these averaged subunit densities. Model-building with an optical comparator and Kendrew parts is in progress.

Organisation of protein shell

The packing of protein subunits in TBSV is schematically illustrated in Fig. 1. The icosahedrally-ordered parts of the virus particle can be described as a compact polyhedral shell, about 30 Å thick, with 90 projections at strict and local twofold positions[3]. This structure is achieved by regular interactions of subunits that fold into two distinct globular domains, connected

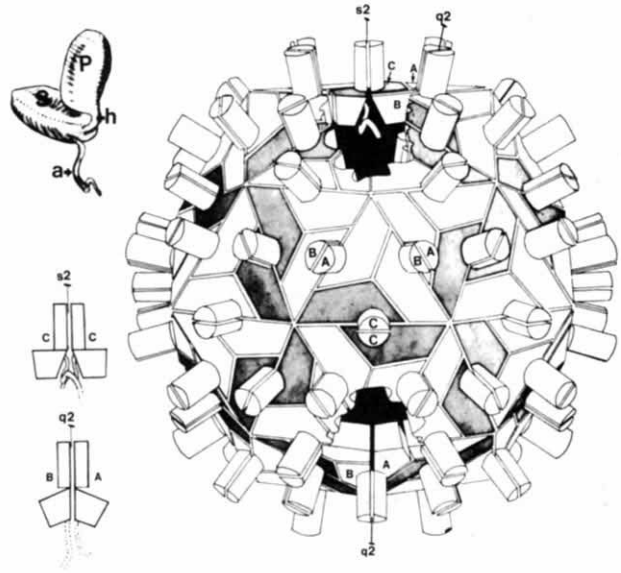


Fig. 1 Packing of protein subunits in TBSV. *a*, Schematic subunit, showing P and S domains, the inter-domain hinge (h) and the N-terminal arm (a). *b*, Arrangement of subunits in the virus particle. A, B and C denote the three packing environments of the subunit; outer surfaces of C subunit S domains are shaded. S domains of A subunits pack around fivefold axes; S domains of B and C alternate around threefolds. The quasi-threefold axis relating S domains of an ABC trimer is nearly parallel to the adjacent strict twofold (s2). Trimers therefore present a rather flat surface across the strict dyad and a distinctly sharper dihedral angle (about 40°) across the quasi dyad (q2). This difference in local curvature can be seen at the two places where the shell has been cut away to reveal S domain packing near strict (top) and quasi (bottom) dyads. *c*, The two states of the TBSV subunit found in this structure, viewed as dimers about strict (s2) and local (q2) twofold axes. Subunits in C positions have the inter-domain hinge 'up' and a cleft between twofold-related S domains into which fold parts of the N-terminal arms. Subunits in the quasi-twofold-related A and B positions have hinge 'down', S domains abutting, and a disordered arm.

by a hinge. We indicate these domains by the letters P (projection) and S (shell)—corresponding respectively to domains 1 and 2 in our previous paper[3]. The 2.9 Å electron density map shows that these domains include only about three-fourths of the sequence, and that the remainder of the chain forms an inward-projecting N-terminal arm. If we denote the three subunit packing modes by A, B and C (compare Fig. 1), then the overall organisation of the virus particle can be summarised as follows. The globular P and S domains fold in an essentially invariant manner. There are two positions of the inter-domain hinge—the A/B configuration (subunits related by a local twofold) and the C configuration (subunits related by a strict twofold), with a relative reorientation of about 20° in passing from one to the other. The A and B backbone conformations and hinge positions are indistinguishable at present, but some differences in side-chain conformations are evident. The conformation of the arm differs radically in position C from its conformation in A or B. Indeed, the map shows no density for the arm on subunits A and B, implying that it is not positionally fixed to the main body of the subunit. On C, however, the arm folds back along the edge of the S domain and interdigitates with two other such arms around the particle threefold axis, forming a remarkable structure described in more detail below. The folding of this part of the subunit depends on its contacts with several other subunits. These interactions seem to lace the particle together in a manner quite different from the cohesion provided by contacts between the independently folded globular domains.

Chain trace

The entire ordered part of the polypeptide chain has been traced without ambiguity (Fig. 3). At a few points, especially near heavy-atom sites, it was helpful to compare subunits in locations A, B and C (because we expect noise and artefact to be different in each), but in general each chain could be followed independently. Polarity was established initially by identification of carbonyls in the one α-helix, and confirmed by the orientation of carbonyls and C_β in β-sheets, by the handedness of β-sheet crossovers and so on.

The N terminus is in the particle interior; the C terminus, on the outside. The sequence of domains along the chain (N-to-C) is therefore: arm, S domain, P domain. The P domain comprises about 110 residues; the S domain, 168. The portion of the N-terminal arm on subunit C that is visible in the map contains

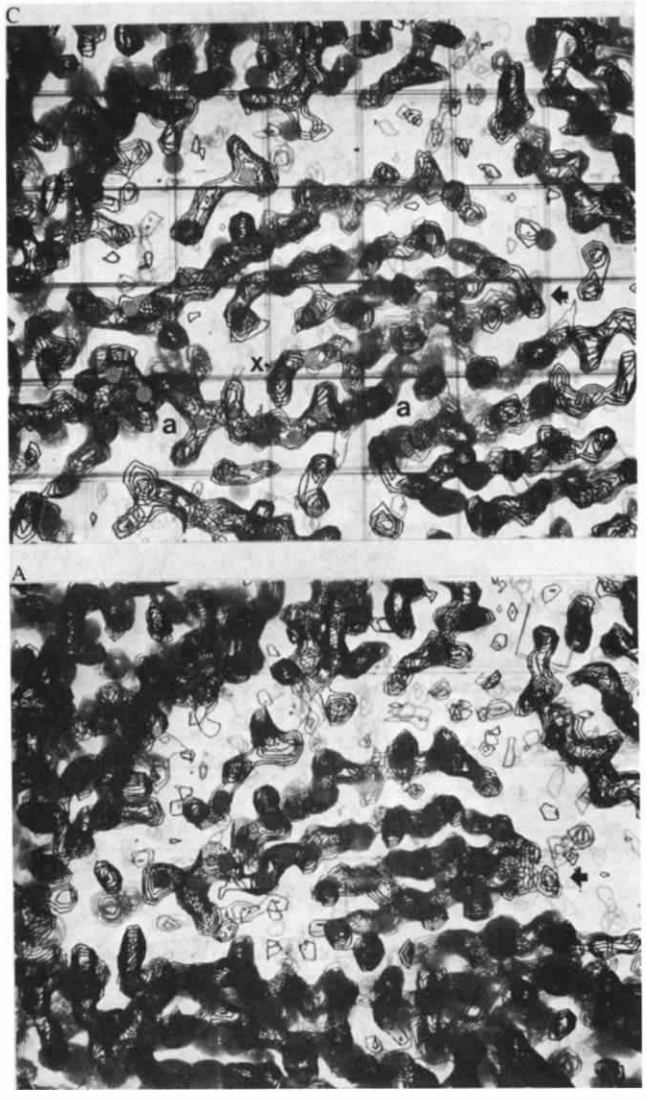


Fig. 2 Portions of the icosahedrally averaged final 2.9 Å electron density map of TBSV. The views are of the inward-facing surface of S domains of subunits A and C, as seen looking down a twofold axis (subunit C, oriented as shaded unit in Fig. 6), or quasi-twofold axis (subunit A). The sectioning of subunit A has been chosen to correspond as closely as possible to the sectioning of subunit C: the invariant fold of the S domain is evident from the comparison. Grid squares correspond to 10 Å. This part of the S domain is a four-stranded β sheet, with the arm (a) folded back alongside it in the case of C. The periphery of each photograph includes features from adjacent subunits, so that this part of the density appears different in each view. Note the large side chain (x), hydrogen-bonded to a backbone carbonyl of the arm of C and disordered on A. Bold arrows show the point where chain folds back into arm on C and extends more irregularly into the centre on A.

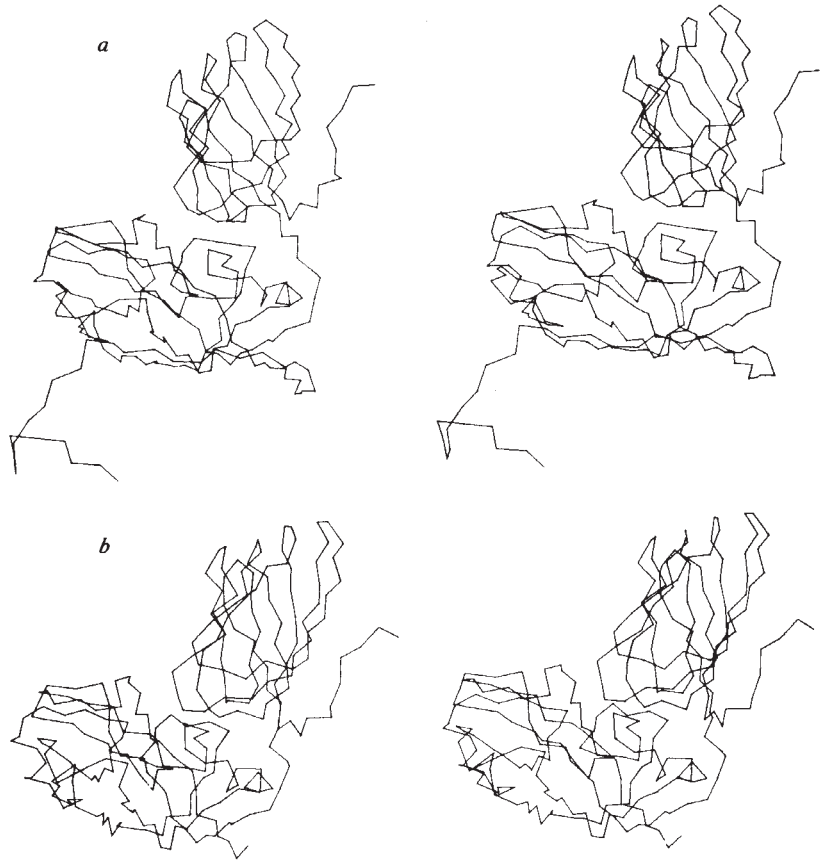


Fig. 3 Stereo views of the two states of the TBSV subunit: (*a*) C, with ordered arm and hinge 'up'; (*b*) A/B, with arm not fixed (and hence not shown) and hinge 'down'. The view of C is normal to the twofold axis, in the direction of the arrow in Fig. 6; the view of A is chosen such that S domains are as closely superposed as possible. The line segments making up these drawings join α-carbon positions, indicated by their intersection. The α-carbon coordinates were measured from mini-maps as described in the text.

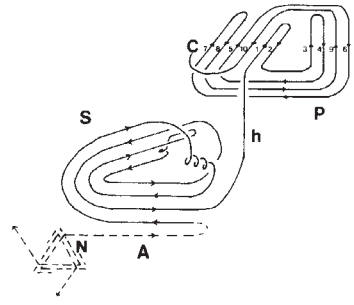
33 residues. The total of 311 identifiable residues corresponds to a molecular weight of ~34,000 (taking 110 as a mean residue weight). As the polypeptide chain molecular weight determined by gel electrophoresis and by gel filtration in guanidine hydrochloride[9,10] is 40,000±2,000, there may be as many as 50 or 60 additional residues, not visible in the map even in subunit C. The connectivity of P and S domains and of the arm is clear. The C terminus also appears to be very clear: moreover, the position of specific sulphydryl cleavages as well as the remarkable protease resistance of the intact particle (A. Andreadis and S.C.H., unpublished) argues strongly against large 'dangling' loops at the outside. The most likely explanation for any difference between the number of visible residues and the expected total is therefore a further segment at the very N terminus. The complete N-terminal arm could then contain as many as 80–90 residues, all of them disordered in positions A and B, and the 33 closest to the S domain ordered in position C. Since the amino acid sequence is unknown, a more exact estimate cannot yet be made. We emphasise that the absence of density corresponding to N-terminal segments does not imply the absence of a regular fold for this part of the polypeptide chain; it requires merely that the terminus not be fixed spatially in the same way on all symmetrically related subunits.

The P domain is composed of two entirely antiparallel β sheets, one of six strands and the other of four (Fig. 4). The six-stranded sheet lies adjacent to a twofold axis and therefore packs against a similar sheet in the related subunit. The curvature of the four-stranded sheet is such that towards the particle centre (bottom of the domain in Fig. 3) it lies at some distance from the six-stranded sheet, with the strands curving gently around from one sheet into the other while preserving a β hydrogen-bond relationship. The two sheets are thus partly continuous along the chain direction (strands 4 to 5, 6 to 7, 8 to 9), although there appears to be no main-chain hydrogen bonding across the ends of the 'sandwich' (that is, between strands 2 and 3 or 6 and 7). About 20 residues at the C terminus form a strand that curls under the domain and up to the terminal residue, which lies at the subunit interface.

The S domain (Fig. 4) is also largely β sheet. It is shaped roughly like a triangular prism; a bent, four-stranded sheet forms the lower surface and one side of the most acute angle of the prism; a second four-stranded sheet forms the other side of the angle. The opposite face, and the upper surface, are formed by loops and by two short helical segments. The longer of these helices (about $2\frac{1}{2}$ turns) is tangentially directed and lies adjacent to a dyad; the other is a quite irregular helix of less than two full turns, radially directed, and packed against a sheet. The loops and helices appear to contain those residues most critically involved in variable inter-subunit or inter-domain contacts (compare 'contacts', below). In subunit C, the N-terminal arm folds back along an inner edge of the prism-shaped domain, forming a fifth element of one of the β sheets.

The inter-domain hinge consists of just two or three residues, with the backbone well ordered in each of the two observed configurations. There are probably small changes in several backbone torsion angles generating the hinge displacement.

Fig. 4 Schematic representation of the TBSV polypeptide chain, showing connectivity of domains and secondary structure within them. Lines with arrows indicate strands of β-pleated sheet (all antiparallel in this case). The N-terminal arm and β-annulus are shown by dashed lines, with segments of the two other chains in the annulus included. P, S, A, h indicate P and S domains, the arm and hinge respectively; N and C mark chain termini.

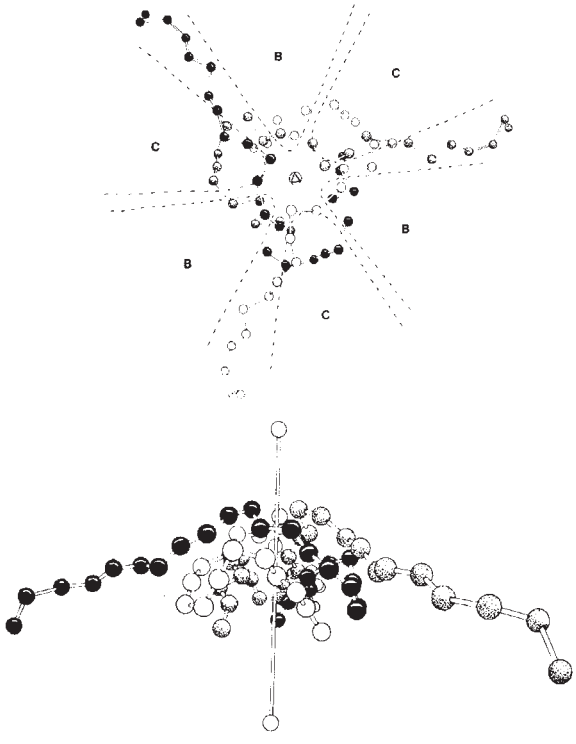


The chain trace described above exhausts all strong features of the map. Additional parts of the polypeptide chains, and all the nucleic acid, must be averaged over a number of orientations with respect to the principal domains. Density ascribed to RNA in interpreting the 5.5 Å map of TBSV ('*e*' in Fig. 2*b* of ref. 3) is now clearly seen to be the arm of subunit C.

N-terminal arm

The most remarkable aspect of the TBSV subunit is the configuration of its N-terminal portion. Following the chain in the C-to-N direction through the S domain in subunits A and B, we observe that the strong, distinct features in the map 'fray-out' and weaken at residue '35', and that any clear chain trace is impossible after about residue '33'. (The numbering scheme adopted here begins at the first visible residue on subunit C. Residue '1' may be up to 50 residues in from the actual N terminus: see preceding section.) By contrast, in subunit C, the chain executes a reverse turn and folds along the inner edge of the S domain next to the twofold axis. Its subsequent course is shown in Fig. 5. Extending towards the particle threefold axis, in the cleft between adjacent S domains, the chain makes a number of contacts with the neighbouring (B-type) subunit. On reaching the region of the threefold axis it winds anticlockwise around the axis (viewed from outside the particle), and inwards, forming a roughly 240° circuit from residues '18' to '1'. As there are, of course, two other symmetry-related chains, the 54 residues form an annular structure around the threefold axis. Their configuration is rather like the overlapping flaps of a cardboard carton: chain 1, over chain 2, over chain 3, over chain 1. This threefold ring seems to be stabilised by several sorts of interactions. Residues 3–5 of one chain, 11–13 of a second and 16–18 of a third form a little three-residue, three-strand β sheet: each chain enters such a sheet three times, once at each level, and the entire structure might well be called a 'β-annulus'. In addition to these main-chain interactions, there are important side-chain contacts. Residues 4 and 15, for example, have side chains that make contact around the threefold, partly defining the 'inside' of this small, globular cluster. Moreover, the entire cluster is suspended inward from the apices of the six neighbouring S domains (three C types, corresponding to the subunits to which the three participating arms 'belong', and three alternately intervening B types).

Fig. 5 Configuration of the β annulus, composed of three interdigitating arms. Balls show α-carbons; sticks are schematic indications of the peptide-chain connection. (*a*), View down threefold axis; (*b*), view normal to threefold axis.



It should be observed that, although the interdigitation of arms in the β annulus is at first sight complex, it presents no problem in principle for folding. No knots are implied, for example, and the structure can form by the interweaving of loops in a manner consistent with the presence of further chain distal to it. Nonetheless, this folding must occur during assembly, since the chain depends for its configuration on interaction with a number of neighbouring subunits. Indeed, the precise final organisation of the arm is clearly possible only when B-type subunits are present as well. In this respect, the arm of the TBSV subunit is similar to the N-terminal arm of LDH[11] and to crystalline glucagon[12], where acquisition of the observed conformation seems to depend on formation of the final assembly.

This map does not show the conformation of the arm of subunits A and B, nor that of any additional subunit C residues beyond the annular structure just described. The absence of strong density corresponding to these portions can have several possible interpretations. The segments in question could adopt more than one orientation from subunit to subunit. The map would then have a 'smeared-out' superposition of these alternative structures. The disorder could also be dynamic with all or part of these segments undergoing rather large-scale diffusional motion. Preliminary NMR measurements (C. Dobson and S.C.H., unpublished) lead us to prefer the former picture.

Subunit contacts

Individual domains are identically folded in all three quasi-equivalent subunit locations. We have not yet analysed intra-domain flexibility in detail, but its magnitude is relatively small (C_α superpositions of individual domains give r.m.s. differences of 1 Å or less, the expected measurement error). Variation is concentrated at the domain interfaces, both within and between subunits, and restricted to alternative positions of side chains rather than of the backbone. We can obtain a first description of the capacity of a TBSV subunit for variable packing by examining the degree of conservation or nonconservation of side chain conformation and interaction at the domain interfaces.

Each local symmetry element gives a class of subunit contact (Fig. 6). These contacts may be further subdivided according to the domains involved. The analysis is summarised in Fig. 6. In general, it seems that contacts are either fully conserved (identical side chain packing at quasi-equivalent positions) or distinctly altered. The fivefold/sixfold contacts are a good illustration. These contacts are 'quasi-equivalent' in the sense that they involve the same parts of the subunit. 'Sixfold' is something of a misnomer, however, except in the sense of coordination. As shown already at 5.5 Å, pairwise contacts of the six S domains packed about a strict threefold axis alternate in character[3]. Proceeding anticlockwise (viewing from outside the particle), C→B contacts are essentially identical to contacts about a fivefold axis, while B→C contacts involve quite different interactions, due to intervention of the C subunit arm. At these, as at other non-conserved interfaces (S domain twofold contacts; P domain/S domain contacts), side chains that make important interactions in one state often appear rather disordered in the other.

The conserved local-threefold interfaces have one particularly interesting feature—an ion, perhaps calcium, coordinated by side chains from both subunits. This site is probably linked to an EDTA-induced swelling observed at *p*H 7.5 (S.C.H. and A.J.O., unpublished).

RNA

There is no strong density in this map that can be ascribed to RNA; the entire inner cavity of the virus particle ($r < 112$ Å) is devoid of dense or interpretable detail. This cavity must, however, contain both the nucleic acid molecule and the protein

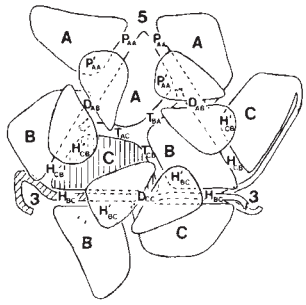


Fig. 6 Schematic diagram showing subunit packing and contacts. The diagram is effectively a more detailed view of the lettered subunits seen nearly face-on in Fig. 1. The domains are indicated in rough outline: barrel-shaped P domains protrude outwards, making only twofold contacts, while the more extensive S domains make contacts across all adjacent symmetry axes. The arms of subunits C fold back as indicated; the arms of A and B are not fixed. The shading of one subunit C indicates which P domain (stippled) is associated with the labelled S domain (hatched). Contacts between subunits are lettered to provide a key for the Table. The cation sites (see text) are near the letters 'T'.

Conservation and shift of inter-subunit contacts

Contact	Code in Figure	Packing
P domain/P domain		
Twofold	D_{CC}	Conserved
Quasi-twofold	D_{AB}	(apposed β-sheets)
S domain/S domain		
Twofold	D_{CC}	Non-conserved
Quasi-twofold	D_{AB}	(Helices roll over each other; sheet edge contacts displaced by arms)
Quasi-threefold	T_{AC}, T_{CB}, T_{BA}	Conserved (Ca^{2+} site)
Fivefold	P_{AA}	Conserved
Quasi-sixfold (fivefold-like)	H_{CB}	
Quasi-sixfold (across cleft)	H_{BC}	Non-conserved (Displaced by arm)
P domain/S domain		
Fivefold	P'_{AA}	Conserved
Quasi-sixfold (fivefold-like)	H'_{CB}	
Quasi-sixfold (across cleft)	H'_{BC}	Non-conserved (Different residue of P domain contacts top of short radial helix)

arms. The volume available to RNA, allowing for the arms, is about 5×10^6 Å^3. This corresponds to about 1.5 g H_2O per g RNA—equivalent to the hydration of DNA in phage heads[13] or to the solvent content of orthorhombic yeast-phenylalanine-tRNA crystals (about 1.3 g per g, ref. 14). TBSV RNA is therefore very tightly packed.

Small angle X ray[15] and neutron scattering from TBSV in solution (B. Jacrot, personal communication) indicate that RNA is present at radii immediately internal to the S domains seen in the present map. We have considered whether dissociation of RNA from fixed binding sites on this domain might have been produced by the relatively high ionic strength (~1.2 M) required for crystallisation. We have succeeded in obtaining high-resolution X-ray photographs from TBSV crystals in 0.1 M $(NH_4)_2SO_4$, 20% polyethylene glycol (ionic strength ~0.3 M): no significant intensity changes are observed at any resolution. Moreover, we note that the inward-facing part of an S domain presents a relatively smooth surface to the interior, unlike the clefts between distinct lobes that characterise the active sites of catenases such as lysozyme and ribonuclease[16]. We conclude that equivalent subunits do not necessarily 'see' RNA in the same position or orientation, thus accounting for the absence of strong high-resolution features.

How, then, does RNA interact with the TBSV subunit? The arguments just outlined imply that the arms of subunits A and B, and the presumed N-terminal part of the arm of C, are interspersed in some way with RNA. Detailed interpretation of neutron scattering suggests considerable penetration of protein towards the particle interior (B. Jacrot, personal communication). We have at present no direct evidence for specific interaction, but the arms are obvious candidates for protein-RNA binding. An RNA-binding region adopting various positions with respect to the body of the subunit might find good contacts, whatever the detailed three-dimensional arrangement of RNA; it would thereby gain the advantage of relative insensitivity to nucleotide sequence.

Implications for assembly

The 2.9 Å map has revealed two quite unexpected aspects of the TBSV structure. A portion of the polypeptide chain near the N terminus interdigitates with others in an unusual fashion in one state of the subunit (C) and appears disordered in the other, quasi-equivalent positions (A/B). Moreover, RNA appears not to bind to the rigidly-anchored S domain; it may instead interact with the flexibly-linked arms. Some interaction of RNA and protein is required for self-assembly.

We have as yet no information about the pathway of TBSV assembly, but we would like to draw attention to one striking feature of the particle design. The threefold-linked arms, together with the D_{cc} contacts, interlace subunits in such a way that all sixty C type subunits are interconnected, without mediation of A or B (see Figs 1, 6). That is, removal of all A and B subunits would leave a still intact, though fenestrated and cage-like, shell. Moreover, all subunits in this shell would be exactly equivalent in conformation and bonding. Twofold (P domain) and threefold (β annulus) contacts in such a structure would determine completely a 60-subunit framework, though perhaps a rather floppy one, connected like the edges of a pentagonal dodecahedron. This hypothetical structure illustrates that no 'switch' is needed to control hinge configuration or size determination. With C subunit arms linked to form a β annulus, the intervening gaps can be occupied only by subunits that adopt an A/B conformation, having hinges 'down' and unlinked N-terminal arms. That is, if linkage of the arms is important in building up a particle, non-equivalent aspects of subunit packing are automatically built into the bonding scheme. Moreover, as A/B subunits are introduced into a network of C's, the structure acquires the rigidity and curvature of a complete T = 3 virus shell.

We are grateful to Dr Carl Moser (and to U., J., and J.W.) for organisation of a CECAM Workshop, at which phase refinement calculations were carried out; to J. Crawford, C. Steele and I. Robinson, for advice and assistance; to Professor W. N. Lipscomb, for use of an Optronics Photoscan; to the Chemistry Department of Harvard University, the Columbia University Computing Center, and the Cambridge University Computing Centre, for computer facilities. The electron density map was plotted using graphics facilities at the MRC Laboratory of Molecular Biology, Cambridge, and we thank C Villa for assistance with the photography. We acknowledge grant support from NIH, a PHS Research Career Development Award and a Sloan Foundation Fellowship (to S.C.H.), a Damon Runyon Fellowship (to A.J.O.), and a Trinity College Research Fellowship (to G.B.). We acknowledge the earlier contribution of the late Dr A. Jack (1948–78).

Received 5 June; accepted 17 August 1978.

1. Caspar, D. L. D. & Klug, A. *Cold Spring Harb. Symp. quant. Biol.* **27**, 1–24 (1962).
2. Bancroft, J. B. *Adv. in Virus Res.* **16**, 99 (1970).
3. Winkler, F. K., Schutt, C. E., Harrison, S. C. & Bricogne, G. *Nature* **265**, 509–513 (1977).
4. Bernal, J. D., Fankuchen, I. & Riley, D. P. *Nature* **142**, 1075 (1938).
5. Caspar, D. L. D. *Nature* **177**, 475–476 (1956).
6. Arndt, U. & Wonnacott, A. *The Rotation Method in Crystallography* (North-Holland, Amsterdam, 1976).
7. Crawford, J. thesis, Harvard Univ. (1977).
8. Bricogne, G. *Acta crystallogr.* A **32**, 832–847 (1976).
9. Ziegler, A., Harrison, S. C. & Leberman, R. *Virology* **59**, 509–515 (1972).
10. Weber, K., Rosenbusch, J. & Harrison, S. C. *Virology* **41**, 763–765 (1970).
11. Adams, M. J. *et al.* in *Protein–Protein Interactions* (eds Jaenicke, R. & Helmreich, E.) (Springer, New York, 1972).
12. Sasaki, K., Dockerill, S., Adamiak, D. A., Tickle, I. & Blundell, T. *Nature* **257**, 751–757 (1975).
13. Earnshaw, W. C. & Harrison, S. C. *Nature* **268**, 598–602 (1977).
14. Kim, S. H. *et al. Science* **185**, 435–440 (1974).
15. Harrison, S. C. *J. molec. Biol.* **42**, 457–483 (1969).
16. Dickerson, R. & Geis, I. *The Structure and Action of Proteins* (Harper and Row, New York, 1969).

letters

Incidence of ionised gas in E/SO radio galaxies

SEVERAL nearby E and SO galaxies are known to contain weak ($L_{rad} \sim 10^{39}$ erg s^{-1}) radio sources[1-3]. Heeschen[4] found that these sources fall into two distinct classes: 'extended' sources with steep radio spectra and sizes comparable to their parent galaxies, and 'compact' sources with flat spectra, which are centred on the optical nucleus and have sizes smaller than a few arc s. Although ionised gas is generally inconspicuous in E/SO galaxies, spectroscopy of ~24 E/SO radio sources[3,5,6] suggests a high incidence of optical emission lines in galaxies with compact radio structure. The two best-known examples, NGC1052 and NGC4278, have, in addition to unusually strong emission lines, abnormal amounts of neutral hydrogen[7-9], and this has led to the speculation[10] that radio outbursts in early type galaxies are fuelled by accretion of neutral gas, possibly from intergalactic clouds. As part of an optical study of E/SO sources with flux densities higher than 20 mJy from the deep Arecibo radio survey of Dressel and Condon[11], we have recently obtained spectrophotometry of 57 objects which makes the radio structure–optical line emission correlation more definitive. The 2,380 MHz Arecibo survey of bright Uppsala General Catalogue galaxies included 2,095 objects with $0° \leq \delta \leq +37°$. 24% of the E, and 12% of the SO galaxies were detected at the 5σ level ($S_{2380} > 15$ mJy). Interferometer observations[12] of the brighter sources allowed separation into the compact (CRS) or extended (ERS) categories; continuing work with higher resolution indicates that several ERS objects have a well-defined double structure (J. J. Condon, personal communication). The optical observations were made with the image dissector scanner on the Kitt Peak National Observatory 2.1 m telescope in December 1977. The scanner was operated with an 8.4-arc s diameter entrance aperture, which yielded an effective resolution of ~11Å. Scans covered λλ 3400–5200. We confine our attention here to [O II] λ 3727, which is expected to be ~10 times stronger against the galaxy background in objects with spectra like NGC1052 (ref. 8) than is Hβ.

The limiting detectibility for our discussion is $W(3727) \sim 0.6$ Å equivalent width. A more refined analysis which includes comparison to line-free absolute energy distributions should yield a limit of ~0.2 Å. For comparison, the limit for [O II] observations by Humason, Mayall and Sandage[13] is ~10 Å. Our observations included all Arecibo E/SO detections with $0\,h < \alpha < 11.7\,h$ or $20\,h < \alpha < 24\,h$ and with $S_{2380} > 30$ mJy, as well as five fainter detections. Additional CRS objects were selected from lists of sources outside the Arecibo declination limits. Finally, a selection of non-radio source (NRS) E/SO galaxies, with $S_{2380} \leq 3$ mJy was included from the Arecibo survey as a control group. Objects with morphological classifications later than SO in the reference catalogue[14] are excluded.

Results are shown in Fig. 1, which plots $W(3727)$ against a radio-optical flux ratio defined as $R = 10^{-5}\, S_{2380}/10^{-0.4m}$, where m is a corrected apparent magnitude described in ref. 12. Apart from the three Seyfert galaxies, two of which (UGC3374 and 5025) are newly identified Seyfert 1 objects, rich emission line spectra were not encountered. (The Seyferts are probably spirals, not E/SO galaxies[15].) Figure 1 also shows a histogram of line strengths in NRS objects. Only 4, or 20%, have detectable λ 3727. For the ERS objects, only 4 out of 16 were detected, and one of these has a compact core component contributing 20% of the total flux. The two strongest ERS [O II] detections (and two of the four NRS cases) exhibit evidence of young stellar populations, and their line emission may be unrelated to their radio properties. In contrast, all 12 Arecibo CRS objects and four out of six non-Arecibo CRS objects were detected. Most of these have $W(3727) > 10$ Å, whereas only one ERS/NRS detection is this strong.

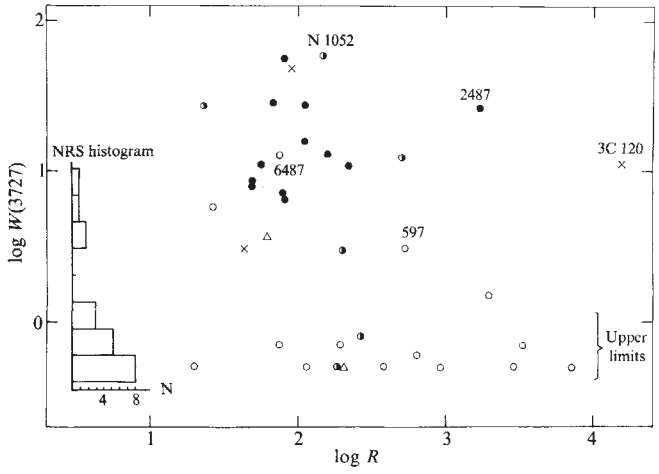


Fig. 1 A plot of the [O II] equivalent width against a radio–optical flux ratio described in the text. Strengths < 1 Å are upper limits. At the left is a histogram of [O II] strengths in non-radio source objects, for all of which $R < 6$. CRS objects have > 50% of their radio flux originating in a region < 6 arc s in diameter. The notation ERS + CRS indicates objects with distinct core components which contribute < 50% of the flux. Labels without prefixes are UGC numbers. ○, ERS; ●, CRS; ◑, CRS, non-Arecibo; △, ERS + CRS; ×, Seyfert.

We thus conclude that among E/SO galaxies, CRS objects have a high probability of strong [O II] line emission whereas ERS objects, despite R values which are typically an order of magnitude higher, are no more likely to exhibit [O II] than galaxies without detectable radio emission. Because most CRS objects are SO galaxies while most ERS objects are E galaxies[12], it is important that this relationship holds for E or SO systems considered separately, although the samples in each category are small. On the basis of these results, one expects that a large fraction of E/SO galaxies with [O II] detected by Humason, Mayall and Sandage[13] will turn out to have compact radio cores. There is no obvious correlation in Fig. 1 between [O II] strengths and the luminosity of the compact source, suggesting that photoionisation by non-thermal radiation may not be responsible for heating the gas.

Because compact radio sources require replenishment of relativistic electrons on a much shorter time scale than extended sources, it is reasonable to expect a correlation like that observed here if infall of thermal gas on to a condensed nuclear object is responsible for the production of relativistic plasma. It now seems that such gas in most cases must originate inside the

0028–0836/78/0276—0374$01.00

© Macmillan Journals Ltd 1978

parent galaxy, as the lack of intergalactic H I clouds[16,17] together with the fact[12] that CRS objects are not preferentially found in galaxy groups or clusters argues that accretion of external gas is rare and probably unnecessary for CRS production. Rather special conditions are required for accretion of hot intracluster gas[18]. Gas lost during normal stellar evolution can readily meet the power requirements of these sources if it is acquired by an accretion disk surrounding a massive black hole.

The high incidence of line emission in CRS objects does not, however, prove that gas infall fuels radio outbursts. The compact source may merely provide a heating mechanism for small amounts of neutral gas (say $\sim 10^6\ M_\odot$, which is impossible to detect with current 21-cm sensitivities[7]) which are incidentally present in E/SO galaxies. In ERS objects, emission might be suppressed through stripping of the thermal gas by an external medium (as most ERS sources apparently reside in clusters or groups), through stripping during the ejection of extended components, or through the strict channelling of relativistic energy in beams which may supply the double sources[19]. Further work on the statistics of the ionised component, its kinematics and its spatial distribution would be of great interest.

ROBERT W. O'CONNELL
LINDA L. DRESSEL

Leander McCormick Observatory,
University of Virginia,
Charlottesville, Virginia 22903

Received 12 August; accepted 2 October 1978.

1. Rogstad, D. H. & Ekers, R. D. *Astrophys. J.* **157**, 481–494 (1969).
2. Heeschen, D. S. *Astr. J.* **75**, 523–529 (1970).
3. Ekers, R. D. & Ekers, J. A. *Astr. Astrophys.* **24**, 247–253 (1973).
4. Heeschen, D. S. *Astrophys. Lett.* **6**, 49–53 (1970).
5. Disney, M. J. & Cromwell, R. H. *Astrophys. J. Lett.* **164**, L35–L39 (1971).
6. Colla, G. *et al. Astr. Astrophys.* **38**, 209–223 (1975).
7. Knapp, G. R., Kerr, F. J. & Williams, B. A. *Astrophys. J.* **222**, 800–814 (1978).
8. Fosbury, R. A. E., Mebold, U., Goss, W. M. & Dopita, M. A. *Mon. Not. R. astr. Soc.* **183**, 549–568 (1978).
9. Knapp, G. R., Gallagher, J. S. & Faber, S. M. *Astr. J.* **83**, 139–146 (1978).
10. Gunn, J. E. *Phys. Scripta* **17**, 277–280 (1978).
11. Dressel, L. L. & Condon, J. J. *Astrophy. J. Suppl.* **36**, 53–75 (1978).
12. Condon, J. J. & Dressel, L. L. *Astrophys. J.* **221**, 456–467 (1978).
13. Humason, M. L., Mayall, N. U. & Sandage, A. R. *Astr. J.* **61**, 97–162 (1956).
14. De Vaucouleurs, G., de Vaucouleurs, A. & Corwin, H. G. *Second Reference Catalogue of Bright Galaxies* (University of Texas Press, Austin, 1976).
15. Adams, T. F. *Astrophys. J. Suppl.* **33**, 19–37 (1977).
16. Haynes, J. P., Brown, R. L. & Roberts, M. S. *Astrophys. J.* **221**, 414–421 (1978).
17. Lo, K. Y. & Sargent, W. L. W. *Astrophys. J.* (in the press).
18. Mathews, W. G. & Bregman, J. N. *Astrophys. J.* **224**, 308–319 (1978).
19. Blandford, R. D. & Rees, M. J. *Phys. Scripta* **17**, 265–274 (1978).

Search for X-ray emission from AO0235+164

THE BL Lacertae object, AO0235+164, is one of the best documented cases of the inverse-Compton scattering problem, that is, the brightness temperature of the emitting region, as implied by rapid variability and cosmological redshift distance, is well in excess of 10^{12} K. If the variable emission is produced by incoherent synchrotron radiation, then at these brightness temperatures the electrons would preferentially Compton scatter from their own synchrotron-emitted photons producing an excessive X-ray flux and subsequent cooling. In late-1975 AO0235+164 underwent an enormous outburst at radio[1,2], infrared and optical[3] wavelengths, for which Ledden *et al.*[2] deduced a brightness temperature $\gtrsim 10^{15}$ K at 8 GHz. This source is also an active variable at low frequencies[4,5] ($\lesssim 1$ GHz). We describe here our search for X-ray emission from AO0235+164.

Explanations of this enormous outburst and related phenomena have invariably involved abandonment of one or more of the basic assumptions[6–9]: incoherent synchrotron emission by electrons, nonrelativistic expansion, cosmological redshift distance, or random geometry. Some of the explanations proffered, such as those involving coherent emission[7], synchrotron proton radiation[6] or special geometry[8], predict that the inferred brightness temperature is the observable brightness temperature of the varying region, although no such high brightness components ($\sim 10^{15-16}$ K) have been detected in recent searches for their expected interstellar scintillations[10–12]. In the case of AO0235+164, unfortunately, such a high brightness component may be unobservable due to possible angular broadening caused by the $z = 0.524$ absorption redshift system.

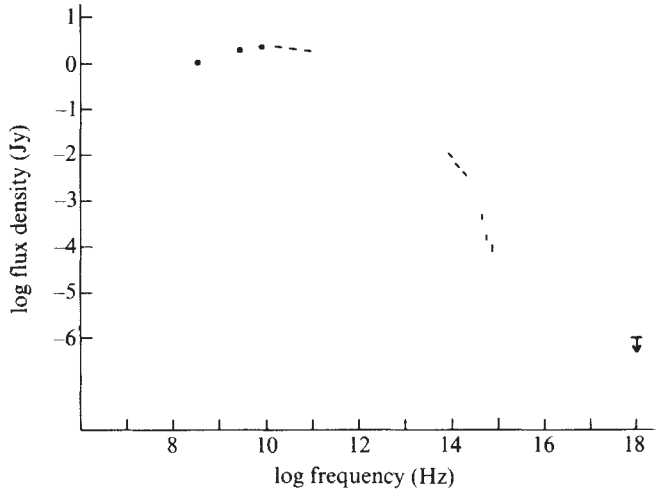


Fig. 1 Spectrum of AO0235+164. Vertical extent of bar (optical data) represents range of galactic extinction corrections used in ref. 21 (December 1977). This data is uncorrected for extinction in the $z = 0.524$ absorption redshift system. Dashed lines are based on observations at other epochs (see text). ●, Data from ref. 5 (March 1978); ↧, this work.

It is clearly important to determine or set restrictive limits to the level of X-ray emission in compact objects. We chose AO0235+164 for our investigation on the basis of its previously discussed properties, and its high density of radiant energy inferred from VLBI observations[4]. Before its late-1975 outburst a limit of $\sim 20\ \mu$Jy at 10^{18} Hz could be set[13]. Since the outburst, the radio flux density has remained well above the 1974 minimum (0.6 Jy) indicating continuing injection of particles (Hugh Aller, personal communication).

The observations were carried out with the rotating modulation collimator (RMC) experiment aboard the SAS 3 X-ray observatory[14,15]. Both RMC systems having FWHM of 2.3 arc min and 4.5 arc min were used in the 2–6 and 6–11 keV bands. AO0235+164 was observed from 17.4 to 22.4 November 1977, a time when the Sun angle for this source was $\sim 180°$, thus maximising the usable data per orbit. The total exposure was 2.6×10^6 cm^2 s. Correlation distributions for different data channels and each orbit were constructed[16] and co-added to produce distributions for the entire observation in both energy bands. No emission was detected in either energy band at or near the best available radio position[17]. The limit to our sensitivity is imposed by counting statistics due to the background. The 3σ upper limit to the counting rate over the entire 2–11 keV band is 0.54 s^{-1}. This corresponds to a total flux limit of 2×10^{-11} ergs s^{-1} cm^{-2} over this range. The flux density at 10^{18} Hz is therefore $\lesssim 1.0\ \mu$Jy. For the relatively flat X-ray spectrum expected for this source, these flux limits are insensitive to the assumed spectral index.

This upper limit allows us to place restrictions on first order Compton scattering, and therefore the brightness temperature and angular size, θ, of the compact radio source[18–20]. The overall spectrum around the time of our observation is shown in Fig. 1. The dashed lines indicate extensions of the radio and optical spectra into the infrared, strongly suggested by observations of this source at other epochs[21]. The optical flux may be depressed by several magnitudes of extinction in the $z = 0.524$ absorption redshift system[4], and thus the B-magnitude does not provide the

0028–0836/78/0276—0375$01.00 © Macmillan Journals Ltd 1978

best limit to first order Compton scattering. At the time of our observation the optical flux was near a minimum[21].

The radio spectrum does not exhibit a well-defined synchrotron self-absorption turnover. The very broad peak at 10 GHz could be best explained by an inhomogeneous synchrotron source[22], each component of which has a fairly flat spectral index. Ledden *et al.*[2] found $\alpha \approx 0.15$ in fitting the outburst spectrum to several self-absorbed components. Conservatively, we shall use $0.2 \leqslant \alpha \leqslant 0.5$, as large values of α produce the smallest lower limits on θ. Apparently, the radio flux decreases beyond 10 GHz. We shall, therefore, consider that the relevant component becomes optically thick at 10 GHz, as this also gives the smallest lower limit on θ. We also assume that the radio spectrum continues to at least 10^{12} Hz before cutting off, as suggested by an apparent continuity between the radio and infrared spectra[21]. This is necessary if Compton scattering is to produce X-rays at $\gamma^2 10^{12}\,\mathrm{Hz} \approx 10^{19}$ Hz.

The maximum brightness temperature, T_n, occurring at frequency ν_n, is given by[20,23]

$$\frac{F^c_{\nu_c}}{F^s_{\nu_s}}\left(\frac{\nu_c}{\nu_s}\right)^{\alpha} \simeq \frac{16\pi}{3}\frac{j^{sc}_{\alpha 0}}{k_{\alpha 0}}\ln\Lambda\frac{r_e\nu_n}{c}\left(\frac{kT_n}{i_{\alpha 0}mc^2}\right)^{3+2\alpha}(1+z)^{4+2\alpha}$$

where $F^s_{\nu_s}$ is the synchrotron flux density at optically thin frequency ν_s, $F^c_{\nu_c}$ is the Compton-scattered flux density at frequency ν_c, $j^{sc}_{\alpha 0}$ and $k_{\alpha 0}$ are dimensionless functions of α (given in ref. 18), $\ln\Lambda$ is a function of the cutoffs in the electron spectrum[18], r_e is the classical electron radius, and m is the electron mass. We then find $T_n \leqslant 2.4\times 10^{11}$ K, and correspondingly $\theta \geqslant 0.38$ m arc s, independent of distance (except through redshift corrections). This limit is derived for $\alpha = 0.5$. If, however, $\alpha = 0.2$, then we obtain $T_n \leqslant 1.5\times 10^{11}$ K and $\theta \geqslant$ 0.50 m arc s.

These values are to be compared with the 10.8 Ghz VLBI angular size measured in January 1978 with a transcontinental baseline (to be reported fully elsewhere by J. Broderick, *et al.*). It was found that $\geqslant$80% of the source is contained within a component of angular size $\leqslant$0.30 m arc s. In view of the various uncertainties involved in the application of the synchro-Compton model, and the interpretation of the VLBI visibility data, we do not consider this to be a major discrepancy. Of interest, however, is the possibility that it may indicate the presence of relativistic expansion in the source. Given the distance implied by the largest absorption redshift present (0.851) we can calculate the physical dimension of the source, l, and therefore the energy contained in relativistic particles, U_e, and magnetic field, U_B (refs 19, 23). Taking $\theta \simeq 0.30$ m arc s and $H_0 = 55\ \mathrm{km\,s^{-1}\,Mpc^{-1}}$, we find that $l \simeq 2.8$ pc, $U_e \simeq 3\times 10^{54}$ erg, and $U_B \simeq 3\times 10^{53}$ erg, for $\alpha \simeq 0.2$. If $\alpha \simeq 0.5$ then $U_e \simeq 1.4\times 10^{55}$ erg and $U_B \simeq 5\times 10^{52}$ erg. In either case the ratio of particle to magnetic energy, E_{em}, is large ($\geqslant$10), and the dynamics must be particle dominated. This would indicate relativistic expansion if the mean energy per particle is comparable to or greater than the rest energy of those particles that neutralise the electron plasma.

If the source is relatively local[6], then the physical size is reduced accordingly, the energy requirements are reduced substantially, but the dynamics remain particle dominated because $E_{em} \propto l^{-1}(1+z)^9$. The minimum of E_{em} occurs for $l \simeq 700$ Mpc, in which case $E_{em} \geqslant 1.5$. Placing the source at a distance at which the brightness temperature inferred from the late-1975 outburst is 10^{12} K ($\sim$200 Mpc) gives $E_{em} \geqslant 3$.

Wolfe *et al.*[4] found that the observed brightness temperature is $\geqslant 10^{12}$ K at $\sim$1 GHz, in excess of the inverse-Compton limit $\simeq 10^{12}\,\mathrm{K}/(1+z) \simeq 5\times 10^{11}$ K. In a static synchro-Compton model this would produce an excessive infrared or optical flux, further requiring that relativistic expansion be invoked.

The magnetic field, **B**, can be calculated independently of distance[19]. If $\theta \simeq 0.30$ m arc s, then $\mathbf{B} \simeq 0.4$–0.12 G depending on assumed spectral index. As long as $\mathbf{B} \leqslant 0.15$ G, then the electrons radiating at 10^{12} Hz are able to cross the source before losing their energy through synchrotron losses.

It would be interesting to determine better the magnitude of the discrepancy reported here. This can be accomplished by observations with the HEAO B X-ray observatory and VLBI observations over a longer baseline. Thus it will be possible to determine the extent to which relativistic expansion or other effects must be included in the synchrotron radiation models.

We thank Dr W. Mayer and F. Pineda for assistance with the observations, and Drs J. J. Broderick and S. L. O'Dell for useful discussions. B.D. was a guest observer for this observation. This work is supported by NASA grant NSG 5221, and partially supported by NASA contract NAS5-11450 and MIT subcontract SC 22152.

BRIAN DENNISON

Department of Physics,
Virginia Polytechnic Institute and State University,
Blacksburg, Virginia 24061

J. P. DELVAILLE
A. EPSTEIN
H. W. SCHNOPPER

Center for Astrophysics,
Harvard College Observatory and Smithsonian Astrophysical Observatory,
Cambridge, Massachusetts 02138

Received 23 August; accepted 28 September 1978.

1. MacLeod, J. M., Andrew, B. H. & Harvey, G. A. *Nature* **260,** 751 (1976).
2. Ledden, J. E., Aller, H. D. & Dent, W. A. *Nature* **260,** 752 (1976).
3. Rieke, G. H. *et al. Nature* **260,** 754 (1976).
4. Wolfe, A. M., Broderick, J. J., Condon, J. J. & Johnston, K. J. *Astrophys. J.* **222,** 752 (1978).
5. Condon, J. J., Ledden, J. E., O'Dell, S. L. & Dennison, B. K. *Astrophys. J.* (submitted).
6. Jones, T. W. & Burbidge, G. R. *Astrophys. J.* **186,** 791 (1973).
7. Petschek, A. G., Colgate, S. A. & Colvin, J. D. *Astrophys. J.* **209,** 356 (1976).
8. Cocke, W. J. & Pacholczyk, A. G. *Nature* **265,** 608 (1977).
9. Marsher, A. P. *Astrophys. J.* **224,** 816 (1978).
10. Scheuer, P. A. G. *Mon. Not. R. astr. Soc.* **177,** 1P (1976).
11. Armstrong, J. W., Spangler, S. R. & Hardee, P. E. *Astr. J.* **82,** 785 (1977).
12. Condon, J. J. & Dennison, B. K. *Astrophys. J.* **224,** 835 (1978).
13. Giacconi, R. *et al. Astrophys. J. Suppl.* **27,** 37 (1974).
14. Schnopper, H. W. *et al. Astrophys. Lett.* **210,** L75 (1976).
15. Schnopper, H. W. *et al. Astrophys. J. Lett.* **215,** L7 (1977).
16. Schnopper, H. W. *et al. Astrophys. J. Lett.* **161,** L161 (1970).
17. Elsmore, B. & Ryle, M. *Mon. Not. R. astr. Soc.* **174,** 411 (1976).
18. Jones, T. W., O'Dell, S. L. & Stein, W. A. *Astrophys.* **188,** 353 (1974).
19. Jones, T. W., O'Dell, S. L. & Stein, W. A. *Astrophys. J.* **192,** 261 (1974).
20. Burbidge, G. R., Jones, T. W. & O'Dell, S. L. *Astrophys. J.* **193,** 43 (1974).
21. O'Dell, S. L. *et al. Astrophys. J.* (in the press).
22. Condon, J. J. & Dressel, L. L. *Astrophys. Lett.* **15,** 203 (1973).
23. Margon, B. *et al. Astrophys.* **207,** 359 (1976).

IUE observations of the eclipsing binary Epsilon Aurigae

THE eclipsing binary Epsilon Aur is a most peculiar binary system and it has not been explained satisfactorily. We report here our observations of this system using the International Ultraviolet Explorer (IUE) collected at the Villafranca Satellite Tracking Station of the European Space Agency (SA).

Epsilon Aur is a single-spectrum binary whose visible component is a supergiant F0 Ia; it is eclipsed once every 27.1 yr by its invisible companion. The eclipse is total and lasts 714 d. The amplitude is 0.8 mag and is independent of wavelength. The eclipse is peculiar for these reasons: (1) during totality the spectrum of the primary F0 Ia star is always seen; (2) from the amplitude of the minimum the two stars might be expected to have about the same magnitude, a point which cannot be reconciled with the fact that only one spectrum is visible at maximum light, even when doubling of lines is expected. These inconsistencies can be removed by assuming that the invisible companion is surrounded by a disk or ring which is the eclipsing body. The fact that the minimum amplitude is independent of the wavelength was explained by assuming that the eclipsing body is a disk or ring of dust particles[1–3] or of ionised atoms

 © Macmillan Journals Ltd 1978

which explain the opacity of the disk or ring as produced by electron scattering only[4].

This second hypothesis was suggested by the characteristics of the spectrum of the eclipsing body, which can be observed during the partial phases of the eclipse[4]. It is not identical to the F0 Ia spectrum, as would be expected if it were due to stellar light simply scattered by dust particles, but the lines originated from non-metastable levels are much fainter than those in the stellar spectrum which indicates that they are formed in a rarefied gas. However, the non-metastable levels of Mg I, Ca I and Fe I are more populated than is expected from the diluted radiation of the primary star at the surface of the eclipsing body. Hence, if the primary were responsible for the excitation of the eclipsing body (as suggested by Kuiper, Struve and Strömgren[5]), it must be richer than expected in UV radiation.

To explain the observations, it has been suggested[4] that the UV radiation does not come from the F0 primary, but from the invisible companion, which is assumed to be a hot star surrounded by a ring of ionised gas. Other suggestions[3,6] were that the secondary could be a black hole; the source of high-temperature radiation required for explaining the spectroscopic features of the eclipsing ring could be provided by the accretion mechanism, heating the gas at $\sim 10^6$ K (ref. 3).

To check these various hypotheses and discover the nature of the companion, we have observed the UV spectrum of Epsilon Aur with IUE, using the low-resolution mode (FWHM 5.5 to 7 Å) for the far UV (λλ 1150–1900 Å), and the high resolution mode (FWHM ~ 0.15 Å) for the near UV (λλ 1900–3200 Å). The observations were made on 19 April 1978 at phase 0.81 P. The exposure times were 30 min in the low-resolution, and 60 min in the high-resolution mode. The low-resolution spectrum indicates the presence of a hot object whose continuum is interrupted by several absorption lines and by a strong emission identified with λ 1302 O I (Fig. 1). The high-resolution spectrum is that expected for an F-type supergiant.

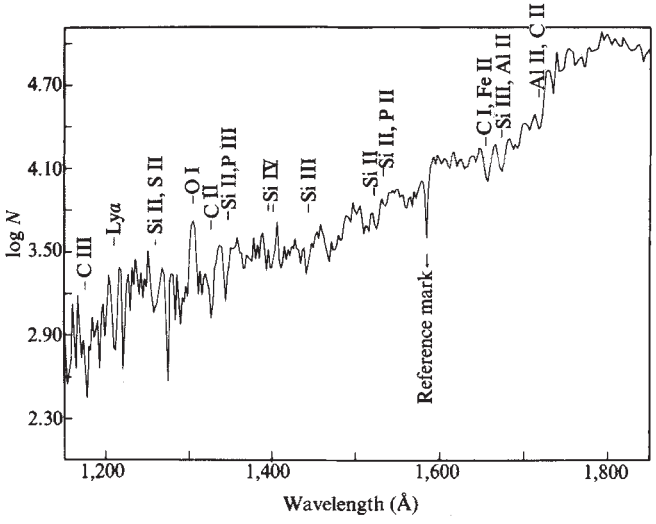


Fig. 1 The low-resolution short-wave spectrum of Epsilon Aur.

The output of the IUE software N, or net flux (corrected for the background), is transformed into absolute flux $\mathscr{F}(\lambda)$ (in erg cm^{-2} s^{-1} t^{-1}) by the relationship $\mathscr{F}(\lambda) = N(\lambda)S(\lambda)^{-1}t^{-1}$ where $S(\lambda)$ is the sensitivity of the IUE camera determined by Bohlin (personal communication) and t the exposure time (s). The values of the sensitivity should be considered as preliminary. Comparison of the fluxes from standard stars observed with OAO 2 and with experiment S2/68 on TD 1 with those observed with IUE (low-resolution mode, large slit 10″×20″) gives an uncertainty of 20% or better (as stated by the resident astronomers). However, our observations were made using the small aperture (3″ hole); in this case the light loss for both spectrographs is estimated to be ~50±20%. The transmission of the small aperture is grey to ±2%. Hence our measured flux is lower than the true flux but the energy distribution is not affected.

The continuum has been traced as a smooth line passing through the highest points of the whole spectrum N versus λ. The faint image at wavelengths shorter than the sensitivity limit at λ 1150 indicates the presence of light scattered from longer wavelengths. The amount of this scattered light is evaluated by measuring the net flux N remaining at the centre of Lyα ($N = 500$). By subtracting this value from the net flux we compute the absolute observed flux corrected for scattered light. Then the flux is corrected for interstellar extinction using the relationship of Nandy *et al.*[7] for $E(B-V) = 0.30$. This value of $E(B-V)$ is obtained by plotting the equivalent width of the interstellar absorption feature centred at λ 2200, $W = 144 \pm 30$ Å, on the relationship W (λ 2200) with $E(B-V)$ given by Nandy *et al.*[7].

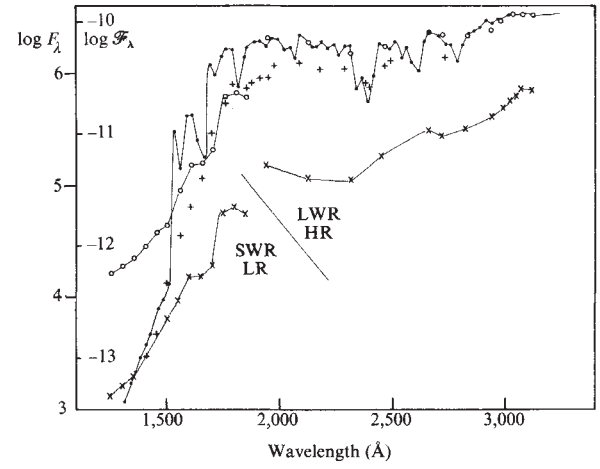


Fig. 2 The continuous flux $\mathscr{F}_\lambda$ (in erg cm^{-2} s^{-1} $Å^{-1}$) observed for Epsilon Aur (×, non-corrected for interstellar extinction, and ○, corrected for interstellar extinction with $E(B-V) = 0.30$), and that observed for Alpha Car with experiment S2/68 on TD 1 divided by 30(+) to make it comparable with that expected for Epsilon Aur which is 3.7 mag less bright than Alpha Car. F_λ is the surface flux computed for an atmosphere with parameters $T_e = 7{,}500$ K and $\log g = 1$ (●). SWP, LR, observations made in the low-resolution mode with the short wave primary camera; LWR, HR, observations made in the high-resolution mode, with the long wave redundant camera.

The continuous flux corrected for scattered light and for interstellar extinction is compared with that observed for Alpha Car (F0 Ia, $E(B-V) = 0.01$) with experiment S2/68 on TD 1 (ref. 8) and with the theoretical surface flux computed by Kurucz et al.[9] for $T_e = 7{,}500$ K and $\log g = 1$, solar composition (a model which reproduces the visual spectrum well[10]) (Fig. 2).

The ratio

$$r_\lambda = \frac{\mathscr{F}_{\lambda\,\mathrm{obs}}}{\mathscr{F}_{\lambda\,\mathrm{pred}}} = \frac{R_A^2 F_{\lambda A} + R_B^2 F_{\lambda B}}{R_A^2 F_{\lambda A}}$$

(where $\mathscr{F}_{\lambda\,\mathrm{obs}}$ is the observed flux at Earth corrected for interstellar extinction, and $\mathscr{F}_{\lambda\,\mathrm{pred}}$ is the predicted flux at Earth computed assuming that Epsilon Aur has the same energy distribution as Alpha Car; $F_{\lambda A}$ is the flux emitted by a stellar atmosphere with $T_e = 7{,}500$ K, $\log g = 1$ computed by Kurucz *et al.*[9]; $R_A = 10^{13}$ cm is the radius of the primary as indicated by the circumstances of the eclipse) has the following values:

$$r(\lambda\ 1300) = 50 \pm 30$$

$$r(\lambda\lambda\ 1400\text{–}1500) = 8.1 \pm 0.4$$

$$r(\lambda\lambda\ 1550\text{–}1600) = 2.3 \pm 0.3$$

$$r(\lambda\lambda\ 1650\text{–}3100) = 1.1 \pm 0.3$$

From these data we can estimate that the effective temperature and the radius of the companion have the following values:

$$T_B = 15{,}000 \pm 3{,}000\ \mathrm{K}, \qquad R_B = 10^{11} \pm 0.5 \times 10^{11}\ \mathrm{cm}$$

The absorption line spectrum confirms that the companion is not very hot: λ 1640 He II and λλ 1548–1550 C IV are not present, while λλ 1394–1402 Si IV are. All the strongest lines of C I, C II, C III, N I, O I (in emission), Si II, Si III, Si IV, P II, P III, S II, S III, and of once- and twice-ionised atoms of the iron group are present.

The strong emission of O I at λ 1302 indicates the existence of an envelope where the hydrogen is mostly ionised, because H I and O I have almost identical ionisation potentials. Lyβ emission can overpopulate the upper level of λ 1025.77 O I, which has almost the same high EP as Lyβ; then by cascade, emissions at λ 8446 and λ 1302 can be explained[11,12].

The companion is therefore a B-type star falling on or slightly below the main sequence, whose absolute bolometric magnitude is about 7 mag fainter than that of the primary. As the system is not close, the presence of the ring cannot be explained by mass exchange between the two stars. The companion may be a past nova, and the ring may have formed with gas expelled during an outburst. Absolute magnitude and spectral type of quiescent recurrent novae have values similar to those derived for Epsilon Aur B.

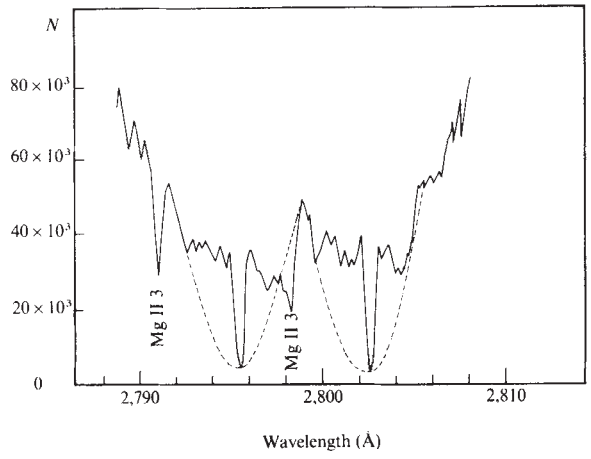


Fig. 3 The profiles of the Mg II resonance lines and the Mg II lines of multiplet 3. The ordinates are IUE counts, corrected for the background. Continuum $N = 130 \times 10^3$.

The near UV spectrum is characterised by strong absorption lines of once-ionised metallic atoms. The most interesting features are the Mg II resonance lines, which have already been observed with BUSS in 1976 but were strongly under-exposed. In the present spectrum (Fig. 3) it is evident that the central part is filled with emission, and this is cut by a strong absorption core which can be formed in the chromosphere of the F0 supergiant or be an interstellar feature, or a blend of the two. From the intensity of the Mg II interstellar lines versus distance relationship given by Lamers and Snijders[13] it seems probable that they are interstellar features. The profile of the Mg II resonance lines is very similar to that observed for Alpha Car[14], but the width of the emissions is 5.1 Å and that of Alpha Car 1.9 Å, suggesting that Epsilon Aur has a larger luminosity than Alpha Car. By plotting the width of the emission w (in km s^{-1}) on the relationship M_v against log w, given by Evans *et al.*[14], a value $M_v = -7$ is found for Epsilon Aur. The flux in the emission is about 2.5×10^{-12} erg cm^{-2} s^{-1} Å^{-1}, and after correction for interstellar extinction is 1.4×10^{-11}. The analogous flux for Alpha Car is 3.1×10^{-10}, that is 22 times that for Epsilon Aur, while the expected value due only to their difference in apparent magnitude is 30 times that of Epsilon Aur.

We thank all the staff of VILSPA Observatory for their assistance in observing and reducing the data, Dr M. Penston for calling our attention to the existence of scattered light in the far UV spectrum, and Dr P. Santin for assistance in data processing with the DEC PDP-11/45 computer system of Trieste Observatory. This work was supported by a CNR-SAS contract.

MARGHERITA HACK

PIER LUIGI SELVELLI*

Astronomical Observatory, Trieste, Italy

Received 22 August; accepted 9 October 1978.

* Present address: VILSPA Observatory, Villafranca del Castillo, Spain.

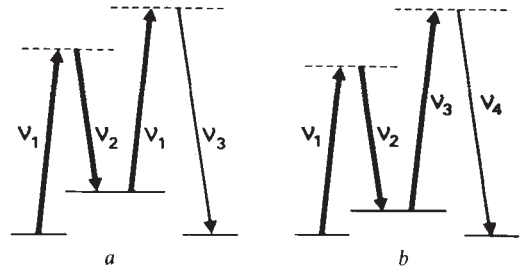
1. Kopal, Z. *Observatory* **74,** 14–20 (1954); *Astrophys. Space Sci.* **10,** 332–339 (1971).
2. Huang, S. S. *Astrophys. J.* **141,** 976–992 (1965); **187,** 87–92 (1974).
3. Wilson, R. E. *Astrophys. J.* **170,** 529–539 (1971).
4. Hack, M. *Astrophys. J.* **129,** 291–313 (1959); *Mem. Soc. Astr. It.* **32,** 351–364 (1961).
5. Kuiper, G. P., Struve, O. & Strömgren, B. *Astrophys. J.* **86,** 570–612 (1937).
6. Cameron, A. G. W. *Nature* **229,** 176–178 (1971).
7. Nandy, K., Thompson, G. I., Jamar, C., Monfils, A. & Wilson, R. *Astr. Astrophys.* **44,** 195–203 (1975).
8. Jamar, C., Macau-Hercot, D., Monfils, A., Thompson, G. I., Houziaux, L. & Wilson, R. *ESA SR-27* (November, 1976).
9. Kurucz, R. L., Peytremann, E. & Avrett, E. H. *Blanketed Model Atmospheres* (published in 1974, new data, unpublished, 1975).
10. Castelli Lanfrè, F. *Astr. Astrophys.* (in the press).
11. Hiltner, W. A. *Astrophys. J.* **105,** 212–214 (1947).
12. Keenan, P. C. & Hynek, J. A. *Astrophys. J.* **111,** 1–10 (1950).
13. Lamers, H. J. G. L. M. ; Snijders, M. A. J. *Astr. Astrophys.* **41,** 259–272 (1975).
14. Evans, R. G., Jordan, C. & Wilson, R. *Mon. Not. R. astr. Soc.* **172,** 585–602 (1975).

Low frequency coherent anti-Stokes Raman spectroscopy of air

THE techniques of coherent Raman spectroscopy are revolutionising the applications and scope of Raman spectroscopy. Particularly important aspects are: very high resolution spectroscopy (the Doppler width being proportional to the Raman shift of the band under study); nanosecond (and sub-nanosecond) spectroscopy, where the resolution obtained may be limited by the Fourier transform of the experimental time scale; discrimination against luminescence from the sample (whether laser induced or not). We discuss here coherent anti-Stokes Raman spectroscopy (CARS) at low frequency shifts.

The technique is briefly as follows: when a sample is irradiated with two lasers at frequencies ν_1 and ν_2 a signal is generated at a frequency $2\nu_1 - \nu_2 = \nu_3$ (see Fig. 1*a*). The signal intensity is proportional to $I^2_{\nu_1} \times I_{\nu_2}$ (where I is the incident beam intensity). When $\nu_1 - \nu_2$ is tuned to be equal to a molecular Raman active transition frequency resonance occurs. Furthermore, the signal at ν_3 is emitted to be a diffraction limited beam, in ideal phase matching conditions. There have been many reports on the use

Fig. 1 Schematic representation of *a*, three-colour CARS; *b*, four-colour CARS.



0028–0836/78/0276—0378$01.00

© Macmillan Journals Ltd 1978

of CARS, notably on gases such as N_2, C_2H_2 (ref. 1) and D_2 (ref. 2). However, we know of only one low-frequency spectrum[3] (equivalent to a Raman shift of ~100 cm^{-1}). In that example the scattering medium was a liquid, while the technique we describe here is suitable for gases at pressures appreciably below atmospheric. Because CARS signals scale normally as the square of the number density, this represents an improvement in sensitivity of more than four orders of magnitude.

The lasers used for our initial experiments were:

Fixed frequency ν_1:	Flash lamp pumped Nd/YAG frequency doubled to 532 nm (JK Lasers)
Tunable frequency ν_2:	Flash lamp pumped dye laser (Chromatix CMX-4).

Using the above equipment, but with the Nd/YAG laser in the unstable resonator configuration, CARS spectroscopy on gases using colinear beams is routine in the range 1,000–3,200 cm^{-1}. Below 1,000 cm^{-1} problems arise due to resonant CARS signals from optical components notably in lenses and windows (where the number density is high). The CARS signal from the glass can be removed conveniently by beam expansion before beam combination and passage through the lenses and windows. Where the cell windows have to be close to the beam waist, CARS signals may be greatly reduced in one of two ways. Either the window material is selected to have no first order Raman spectrum and contains atoms of low polarisability (such as LiF) or the windows are very thin. In this way spectra may be run relatively routinely to 600 cm^{-1} for moderately scattering gases at tens of torr pressure.

Below 600 cm^{-1} the major problem lies in the collimated broad band superradiance emitted at the signal frequency by the tunable dye laser. An additional difficulty is the separation of the weak (wanted) signal from the other two intense (unwanted) colinear beams. Both of these problems become increasingly important as one moves to smaller and smaller Raman shifts.

An alternative experiment uses two tunable dye lasers (ν_1, ν_2) and one fixed frequency laser (ν_3) which is considerably removed in frequency from both the dye lasers, as shown in Fig. 1*b*. Providing ν_1 and ν_2 are well removed from ν_3 in frequency, the dye laser beams can be readily separated from the fixed frequency laser ν_3 with its associated CARS signal ν_4. Superradiance problems with fixed frequency solid state lasers are negligible compared with those from dye lasers.

This arrangement was realised experimentally with a purpose built step-tunable dye laser (ν_1) pumped by 80% of the frequency doubled Nd/YAG output. The remaining 20% of this 532 nm radiation was used as the fixed frequency (ν_3), while the Chromatix CMX-4 acted as the tunable laser (ν_2). In this way, ν_3 is particularly free from other frequencies due to the square law dependence (on the intensity of the incident beam) for the frequency doubling process.

The remaining difficulty is the separation of ν_3 from the signal. This may be achieved by using prisms or gratings of sufficient dispersion. However, for pure rotational Raman spectroscopy all bands are depolarised. In three-colour CARS (Fig. 1*a*) the polarisation of the CARS radiation follows that of ν_2. The situation is more complex in four-colour CARS[4] (Fig. 1*b*). For depolarised bands (or polarised bands where the anisotropy is large) polarisation of the CARS signal (ν_4) will be exactly perpendicular to ν_3 if the polarisations of ν_1 and ν_2 are orthogonal and ν_3 is parallel to either ν_1 or ν_2. By passing all four beams through an analysing polariser and a prism of moderate dispersion, ν_1 and ν_2 can be effectively removed and ν_3 attenuated by six orders of magnitude. The limitation in the removal of ν_3 is the birefringence of the optical components. Additional filtering of the signal with a conventional double monochromator allows the monitoring of ν_4 to approximately a shift of 40 cm^{-1}.

The success of this approach is clearly demonstrated by the pure rotational CARS spectrum of air in Fig. 2. The inset shows

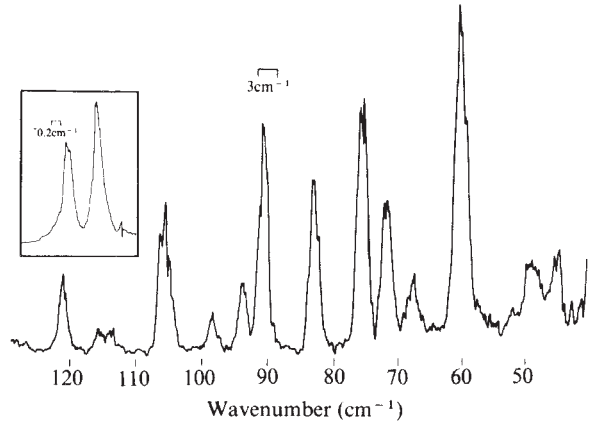


Fig. 2 Pure rotational CARS spectrum of air. Inset: the 60 cm^{-1} band under higher resolution.

a higher resolution spectrum of the 60 cm^{-1} band arising from a near coincidence (~0.7 cm^{-1}) of $^{14}N_2 J = 6 \rightarrow 8$ and $^{16}O_2$ $J = 9 \rightarrow 11$. The resolution of 0.2 cm^{-1} is routine and 0.01 cm^{-1} could readily be achieved, although not using the CMX-4.

For strongly polarised bands the BOXCARS approach of Eckbreth[5], using non colinear beams, coupled with the present four colour system seems to be attractive.

We thank the SRC for financial support, Dr D. C. Hanna and L. Laycock for the design and construction of a Nd/YAG pumped dye laser, and Dr M. Yuratich for helpful discussions.

I. R. BEATTIE
T. R. GILSON
D. A. GREENHALGH

Department of Chemistry,
The University,
Southampton, UK

Received 21 August; accepted 11 October 1978.

1. Nitsch, W. & Kiefer, W. *Opt. Commun.* **23,** 240 (1978).
2. Nibler, J. W., McDonald, J. R. & Harvey, A. B. *Opt. Commun.* **18,** 371 (1976).
3. Beattie, I. R. & Black, J. D. *Vibrational Spectroscopy—Modern Trends* (eds Barnes, A. J. & Orville-Thomas, W. J.) 21–31 (Elsevier, Amsterdam, 1977).
4. Beattie, I. R., Greenhalgh, D. A. & Yuratich, M. (in preparation).
5. Eckbreth, A. C. *Appl. phys. Lett.* **32,** 421 (1978).

Trace elements in diamonds of different types

STUDIES of optical and electrical properties have shown that natural diamond can be broadly categorised into four types, types IA, IB, IIA and IIB (refs 1–3). Type IA is the most common (98% abundance)[3] and contains up to 0.34% N (ref. 16), which has been argued[5] to occur in atom pairs. Type IB diamond contains much less N, of which 1–2% occurs as single substitutional paramagnetic N. Type II diamonds contain even less N. The p-type semiconducting property of type IIB diamond is due to the presence of B (refs 6, 7), the concentration of which has been correlated with the acceptor concentration[7]. We have determined and report here the impurity contents of 23 diamonds of different types in order to investigate whether elements other than N and B contribute to the physical properties of diamond. Although there are few diamonds of each type and the statistics are therefore poor, two diamonds were found to contain an excess of Ni, while the three Type IIB diamonds possibly contained an excess of Al.

The diamonds analysed were classified into the different types using infrared and ultraviolet spectral characteristics, and

0028–0836/78/0276—0379$01.00 © Macmillan Journals Ltd 1978

conductivity measurements. Diamonds 1 to 6 (Table 1) were flawless light yellow octahedra. In all the other diamonds inclusions and flaws were visible at 50× magnification under polarised light and these were removed by cutting and polishing. The individual diamonds were analysed using an established neutron activation analysis technique[4,8]. Superficial contamination was avoided by chemical etching after irradiation to remove a few micrometres of diamond from the outer surface[9]. The results are given in Table 1.

Errors vary with sample size and impurity concentration, and range from 5% at best to 40% at the limit of detection. The most reliable data are for Na, Al, Sc, Mn, Co and Ni. Because of the interest in Ni, limits of detection were calculated for this element[10]. Rare earth elements (REEs) were observed in some diamonds at very low concentrations, but no reliable quantitative determinations could be made.

Despite the care taken to remove inclusions it is clear from the trace element concentrations and their inter-relationships[8,11] that diamonds nos 4, 8, 9, 10, 18 and 20 contained mineral inclusions (Table 1). The impurities present in the rest of the diamonds were inferred to be predominantly due to submicroscopic inclusions of the parental magma[11].

Considering, for example, the relationship between Al and Mn for the 23 diamonds studied (Fig. 1), and ignoring the inclusion-containing stones, there seems to be no significant difference between types IA and IIA diamonds. This observation is supported by more detailed scrutiny of the data in Table 1, and it seems that the present data on trace element levels can contribute little to the existing body of knowledge which attributes the difference between types IA and IIA diamond to the N content, and the physical mode of its occurrence.

The data in Table 1 show that among the inclusion-free diamonds the type IIB diamonds contained the highest levels of Al. A plot of the Al against Mn contents (Fig. 1) shows that the Al content is in excess of that expected from the submicroscopic inclusions of parental magna inferred to be present in all diamonds[11]. That this is an Al excess and not a Mn deficiency can be shown by comparing the Al with other lithophile elements, for example, the Al to Sc ratio is ~3,000 in the parental magma (H.W.F. unpublished data). It is difficult to determine the real significance of the Al excess in type IIB diamonds because comparable Al excesses are occasionally found in other types of diamond, for example, nos 2, 3 and 7 (Table 1). There has been controversy over the part played by Al in semiconducting diamond[12], and it has been suggested that Al

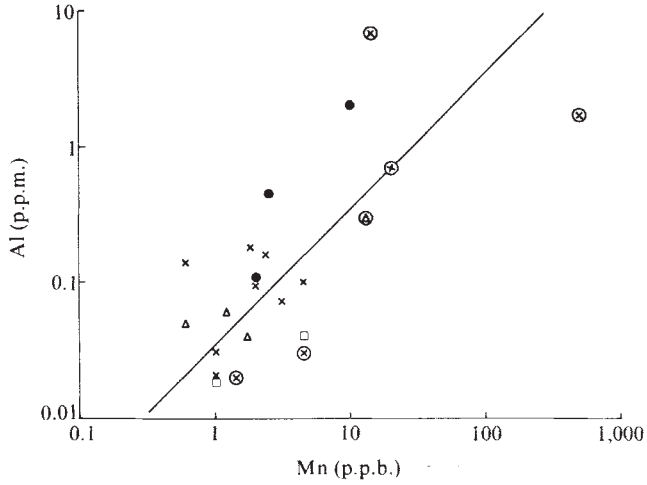


Fig. 1 Aluminium and manganese content in diamonds of different types: ×, type IA; □, type IA; △, type IIIA; ●, type IIB. Points circled are from diamonds containing inferred mineral inclusions. The composition line shows the Al:Mn ratio of ~35 inferred by Fesq *et al.*[11] for the parental magma (p.p.b., parts per 10^9).

Table 1 Trace element concentrations in typed stones

Diamond	Mass (mg)	Na (p.p.b.)*	Mg (p.p.m.)	Al (p.p.m.)	K (p.p.m.)	Sc (p.p.b.)*	Ti (p.p.m.)	V (p.p.b.)*	Cr (p.p.b.)*	Mn (p.p.b.)*	Fe (p.p.m.)	Co (p.p.b.)*	Ni (p.p.m.)	As (p.p.b.)*
Type IA														
1	76	0.1	—	—	—	0.01	—	—	—	0.2	1.0	2.3	0.07	0.01
2	68	0.1	—	0.14	—	0.02	0.05	—	—	0.6	0.3	0.4	0.01	—
3	57	—	—	0.16	—	0.02	0.1	—	—	2.4	—	0.3	<0.01	—
4†	54	1.0	—	0.02	—	—	—	0.6	—	1.5	0.3	4.5	0.07	—
5	59	1.0	—	0.03	—	0.02	—	0.7	—	1.0	—	0.3	<0.01	0.02
6	38	—	—	0.02	—	0.04	0.05	—	—	1.0	0.2	0.3	<0.012	—
7	98	0.1	—	0.18	—	0.01	—	—	0.5	1.8	—	1.0	<0.015	—
8†	40	2900	8.0	6.9	0.17	0.6	—	20	110	14	1.3	3.0	0.08	—
9†	32	260	200	1.7	0.07	2.2	—	15	950	500	47	200	7.0	—
10†	38	190	0.6	0.03	0.03	0.04	—	0.8	2.0	4.5	1.0	0.7	<0.012	—
16	25	6.0	—	0.10	—	0.02	—	1.2	—	4.3	0.2	0.6	<0.015	—
17	27	0.4	—	0.1	—	0.04	—	—	—	2.0	0.3	0.4	<0.015	—
18†	80	0.8	—	0.7	—	—	0.25	3.2	90	20	30	340	40	0.6
19	16	—	—	0.07	—	0.08	—	—	—	3.0	0.2	0.7	<0.03	0.03
20†	28	35	6.0	0.3	0.01	0.03	—	—	2.0	13	0.6	2.0	0.07	—
Type IB														
11	132	0.2	—	0.02	—	0.01	—	—	—	1.0	<0.05	0.4	1.0	—
12	50	0.1	—	0.04	—	0.03	—	—	—	4.5	<0.1	0.5	0.7	—
Type IIA														
13	128	2.0	—	0.04	—	0.01	—	—	0.8	1.7	—	0.1	<0.005	—
14	42	—	—	0.06	—	0.03	—	—	—	1.2	0.1	0.3	<0.016	—
15	116	0.6	—	0.05	—	0.01	—	0.3	—	0.6	—	0.1	<0.004	0.01
Type IIB														
21	105	1.0	1.0	2.0	—	0.1	0.25	—	1.0	10	0.1	0.4	<0.006	0.01
22	44	1.0	—	0.44	—	0.03	0.35	—	—	2.5	0.1	0.3	<0.012	—
23	107	0.2	—	0.1	—	0.02	0.08	—	—	2.0	—	0.1	<0.004	—

Some elements were determined in only a few of the diamonds. Diamond no. 8 contained 20 p.p.m. Ca, Diamonds nos 9 and 18 contained 0.5 and 1.1 p.p.m. Cu, respectively. REE's were observed in diamonds nos 4, 8, 10, 18 and 20.

*Parts per 10^9.

†Diamonds with inferred mineral inclusions.

could act as a getter for N (ref. 13), thus preventing compensation of B. A larger number of type IIB diamonds is being analysed for B, N and Al in order to clarify any secondary role which may be played by Al in semiconducting diamonds.

As Table 1 shows, two diamonds with types IB character (Nos 11 and 12) have unusually large Ni:Co ratios (2,500 and 1,400) compared to all other diamonds. With the exception of these two diamonds, the Ni:Co ratios in the diamonds analysed here ranged from <15 to 118. This range includes the ratios for those diamonds where Ni was not detected but was assumed to be present at the limit of detection. This range is in good agreement with that found previously where Ni:Co ratios ranging from 5 to 110 were found in a large suite of diamonds[4] (note error in ref. 11, Table 3). Those diamonds were not typed but probably contained predominantly type IA diamonds, the most common type. One exception found in that work was for a 1 g sample containing about 15 orange diamonds from the Premier Mine, Transvaal. This sample, which probably contained some type IB diamonds, had a Ni:Co ratio of 250, which again is unusually high. These ratios should be compared to those inferred for the parental magma: silicate phase ~15; sulphide phase ~40 (H.W.F. unpublished data).

In this case it is more difficult to prove which of the two elements is anomalous. However, it is unusual for the Fe:Ni ratio to be <1. In the data obtained from the analysis of over 100 samples of diamond only nine cases were found where the Fe:Ni ratio was <1, the lowest being 0.14. From the calculated limits of detection for iron in diamonds nos 11 and 12, the maximum possible ratios are seen to be 0.05 and 0.14, respectively, which are unusually low[4]. Because both the Fe and the Co contents are low compared to that of Ni in diamonds nos 11 and 12 only, it suggests that these diamonds are characterised by a net Ni excess and are therefore unusual.

Two arguments can explain the high Ni:Co ratio found in these two diamonds. First, it could be due to submicroscopic inclusions of a Ni-rich, Co-poor, phase. The low levels of lithophile elements (for example, Al and Sc) found in these two diamonds suggests that any possible Ni-rich phase should be a sulphide. Alternatively, it has been shown[14] that Ni can occupy lattice sites in synthetic diamonds grown from Ni-rich melts, where Ni acts as a catalyst. Klyuev *et al.*[15] regard the morphology of type IB diamond, and the presence of paramagnetic nitrogen, as indicating formation in a low-temperature environment. If this is so then the availability of Ni as a catalyst could be essential to their formation. Unfortunately the two Ni-rich stones cannot be unambiguously classified: diamond no. 12 is dominantly type IB with some admixture, while diamond no. 11 is most likely mixed type IB and type HA.

These speculations reflect the need for further measurement of the Ni:Co ratios in type IB diamonds. High purity diamonds are necessary as the presence of inclusions with a 'normal' silicate or sulphide composition will tend to mask any unusual Ni:Co ratio. A search for Ni in the lattice of type IB diamonds using other methods of analysis could help clarify the possible role of Ni in the formation of type IB diamonds.

We thank the de Beers Diamond Research Laboratory, the South African Atomic Energy Board, the National Institute for Metallurgy, the University of the Witwatersrand and Drs B. D. Bibby, C. S. Erasmus and E. J. D. Kable for support and assistance.

D. M. BIBBY*
H. W. FESQ
J. P. F. SELLSCHOP†

Nuclear Physics Research Unit, University of the Witwatersrand, Johannesburg, South Africa

Received 6 March; accepted 28 March 1978.

*Present address: Chemistry Division, D.S.I.R., Private Bag, Petone, New Zealand.
†On secondment from the South African Atomic Energy Board.

1. Robertson, R., Fox, J. F. & Martin, A. E. *Phil. Trans. R. Soc.* **A232**, 463–535 (1934).
2. Custers, J. F. H. *Physica* **18**, 489–496 (1952).
3. Dyer, H. B., Raal, F. A., Du Preez, L. & Loubser, J. H. N. *Phil. Mag.* **11**, 763–774 (1965).
4. Bibby, D. M., Erasmus, C. S., Fesq, H. W. & Kable, E. J. D. *Rep. 1638* (National Inst. Metallurgy, Johannesburg, 1975).
5. Davis, G. *J. Phys. C: Solid St. Phys.* **9**, L537–542 (1976).
6. Sellschop, J. P. F., Bibby, D. M., Keddy, R. J., Mingay, D. W. & Renan, M. J. *Proc. Diamond Conf.* Cambridge, (1975); *Int. J. Radiat. Isot.* **28**, 277–279 (1977).
7. Lightowlers, E. C. & Collins, A. T. *J. Phys. D: Appl. Phys.* **9**, 951–963 (1976).
8. Fesq, H. W., Bibby, D. M., Sellschop, J. P. F. & Watterson, J. I. W. *J. Rádioanalyt. Chem.* **17**, 195–216 (1973).
9. Bibby, D. M. *Int. Jl. appl. Radiat. Isotopes*, **28**, 331–337 (1977).
10. Lukens, H. R. in *Advances in Activation Analysis* (Academic, London, 1972).
11. Fesq, H. W., Bibby, D. M., Erasmus, C. S., Kable, E. J. D. & Sellschop, J. P. F. *Phys. Chem. Earth* **9**, 817–836 (1975).
12. Collins, A. T. & Williams, A. W. S. *J. Phys. C* **4**, 1789–1800 (1971).
13. Strong, H. M. & Chrenko, R. M. *J. Phys. Chem.* **75**, 1838–1843 (1971).
14. Samoilovich, M. I., Bezrukov, G. N. & Butuzov, V. P. *J. exp. theor. Phys. Lett.* **14**, 379–381 (1971).
15. Klyuev, Yu. A., Dudenkov, Yu. A. & Nepsha, V. I. *Geokhimya* **7**, 1029–1036 (1973).
16. Sellschop, J. P. F., Annegarn, H. J., Keddy, R. J., Madiba, C. C. P. & Renan, M. J. *Nucl. Instrum. Meth.* **149**, 321–328 (1978).

A reduction in $^{87}Sr/^{86}Sr$ during basalt alteration

HYDROTHERMAL alteration of basalts on the ocean floor has been reported to cause a general increase in the $^{87}Sr/^{86}Sr$ ratio of the basalts[1,2]. Results are presented here from Mull in the British Tertiary province which suggest strongly that hydrothermal alteration in such discrete magmatic centres may actually cause a reduction in the $^{87}Sr/^{86}Sr$ ratios of some basalts. Preliminary data indicate that the $^{143}Nd/^{144}Nd$ ratios are unaffected.

Oxygen isotope studies[3,4] of Tertiary igneous rocks from Skye, Mull and Ardnamurchan have demonstrated that large scale hydrothermal systems were established in these areas at the time of igneous intrusion. On both Skye and Mull hydrothermal aureoles extend for up to 6 km from the central intrusive complexes into the surrounding basalts. Within these regions greenschist-facies mineral assemblages have been formed and the rocks are depleted in ^{18}O due to interaction with hot meteoric waters.

The samples selected for this study are from a quarry 3.5 km east of Salen, which lies within the epidote zone of the central aureole on Mull. A hydrothermally altered lava flow is exposed which shows a gradual transition from dark basalt in the massive flow interior to green 'spilitic' material at the margin. In the centre of the flow only olivine has broken down; it is pseudomorphed by chlorite and oxides and the groundmass is rich in chlorite. With increasing alteration plagioclase and the titanomagnetites are replaced by albite, sphene, and calcite and/or epidote. Only at the margin of the flow is pyroxene completely broken down (Fig. 1) and an assemblage of chlorite, albite, quartz, amphibole, epidote, and sphene occurs in which none of the original mineralogy is preserved. Within this flow Al, Ti, Zr, Nb, Y, P, and the rare earth elements preserve almost constant ratios except at the margin where slight loss of *P* and the light rare earths (La, Ce) is observed[5–7]. This is accompanied by variations in the other major and trace elements; particularly, drastic leaching of Ca. Sr isotope analyses were carried out on 10 whole rock samples across the lava flow and on leached pyroxene separated from the flow centre. A secondary mineral pod consisting of zeolite, calcite, and albite surrounded by chlorite which formed immediately above the base of the overlying lava flow, and two discordant veins were also analysed to determine the $^{87}Sr/^{86}Sr$ ratio of at least some of the hydrothermal fluids.

The results are presented in Table 1. Within the lava flow the age-corrected initial $^{87}Sr/^{86}Sr$ ratios range from 0.70467 near the top of the flow to 0.70519 at the base of the sampled section (Fig. 1). The decrease in $^{87}Sr/^{86}Sr$ towards the top of the flow is accompanied by an increase in the degree of visible alteration and a decrease in Fe^{3+}/Fe^{2+}. Sr contents range between 270 and 440 p.p.m. in the more altered samples but seem to be much less variable near the centre of the flow.

$^{143}Nd/^{144}Nd$ ratios were determined on two samples using the technique described in ref. 8. The two samples show very

0028–0836/78/0276—0381$01.00

© Macmillan Journals Ltd 1978

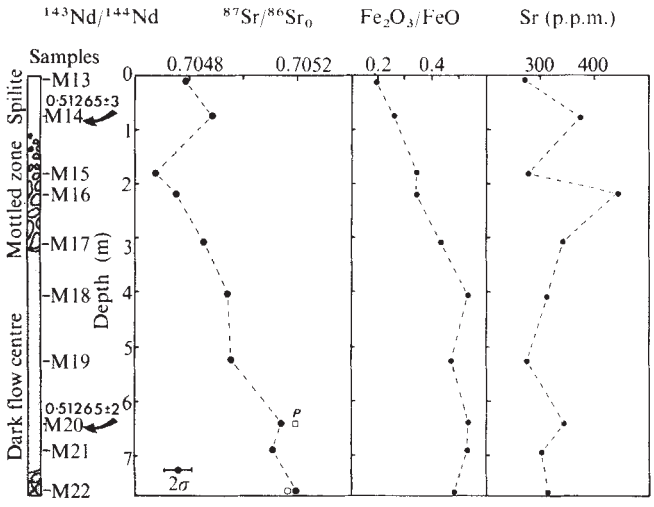


Fig. 1 Variations in $^{87}Sr/^{86}Sr_0$, Fe_2O_3/FeO, and Sr p.p.m. across 7.5 m of an individual lava flow, near Salen, Mull. The degree of hydrothermal alteration is illustrated and M22 shows the onset of increased alteration approaching the base of the flow. Open symbols denote samples leached in 6M HCl; □, pyroxene; ○, whole rock.

different degrees of alteration but in sharp contrast to their variable $^{87}Sr/^{86}Sr$ ratios (Fig. 1) their Nd compositions are indistinguishable analytically (M14, 0.51265±3; M20, 0.51265±2; 2σ errors).

To investigate further the Sr isotope geochemistry, pyroxene was separated from M20 and it and some whole rock powders were subjected to leaching experiments. Leaching was carried out with 6M HCl in sealed Teflon bombs for 14 h at 130°C (ref. 9). X-ray analysis of the whole rock powders before and after leaching showed a marked increase in the concentration of pyroxene and to a lesser extent plagioclase, due to the dissolution of secondary minerals such as chlorite. Analysis of the residue after leaching should therefore provide $^{87}Sr/^{86}Sr$ ratios close to the pre-alteration composition of the lava flow.

The pyroxene has an age-corrected initial $^{87}Sr/^{86}Sr$ ratio of 0.70519 while that of the leached whole rock sample M22 is 0.70516 (Fig. 1). Rb and Sr concentrations for the leached residues from M18, M19, and M20 are not available but they all have significantly higher present-day $^{87}Sr/^{86}Sr$ ratios than the unleached samples. This indicates that the altered material tends to have lower $^{87}Sr/^{86}Sr$ ratios than the fresh basalts.

The leaching experiments and the observed decrease of $^{87}Sr/^{86}Sr$ with increasing bulk rock alteration and decreasing Fe^{3+}/Fe^{2+} (Fig. 1), all suggest that hydrothermal alteration of this lava flow resulted in a general reduction in $^{87}Sr/^{86}Sr$. The composition of the separated leached pyroxene provides the best estimate of the primary composition of the magma (~0.7052). The two $^{143}Nd/^{144}Nd$ results, however, indicate that the lava was isotopically homogeneous when it crystallised, and that Nd was not affected significantly by the hydrothermal alteration (see also ref. 7).

Evidence that the fluids responsible for the hydrothermal alteration did have low $^{87}Sr/^{86}Sr$ ratios is provided by the results on secondary minerals (Table 1). The secondary pod from the overlying flow gave values of 0.70421 and 0.70427. The two vein samples M64 and M24 which cut the flow and therefore represent later stage fluids have $^{87}Sr/^{86}Sr$ ratios of 0.70467 and 0.70521. This may also indicate a rise in the value of the $^{87}Sr/^{86}Sr$ ratio of the fluids with time at this locality.

Forester and Taylor[4] have suggested that the alteration within the hydrothermal aureole on Mull was caused by interaction with fluids having $\delta^{18}O$ values of about −11 to −12. Such depleted oxygen compositions are suggestive of meteoric water in high latitudes, but at least in the area of this study the fluids are also apparently characterised by low $^{87}Sr/^{86}Sr$ ratios (< 0.7048). This clearly rules out interaction with Moinian basement[10], or Tertiary intrusives such as the Glen Cannel and Loch Usig granophyres which have initial $^{87}Sr/^{86}Sr$ ratios of 0.7094±3 and 0.7135–0.7162 (refs 11 and 10) respectively. Sources of the unradiogenic Sr in the fluids are either unconfirmed but possible granulite facies basement, or, much more likely, the basalts themselves.

We envisage that interaction of meteoric water with hot volcanic rocks caused the chemistry of the resultant fluids to be buffered by that of the basalts. As 70% of the fresh volcanic rocks from Mull analysed by Beckinsale *et al.*[11] have $^{87}Sr/^{86}Sr$ ratios <0.7043, the fluids in this area are likely to have similar $^{87}Sr/^{86}Sr$ ratios. Their interaction with basalt of Sr composition 0.7052 will therefore result in a reduction in $^{87}Sr/^{86}Sr$ in the

Table 1 Rb–Sr results

Sample	Lithology/mineralogy	Rb*	Sr*	Rb/Sr	$^{87}Sr/^{86}Sr$†	$(^{87}Sr/^{86}Sr)_0$‡
M13	Altered basalt, chl–plag–qtz–sph	7.17	271	0.026	0.70486±3	0.70479
M14	Altered basalt, some relict cpx	1.76	375	0.005	0.70489±3	0.70488
M15	Altered basalt, increasing growth of	2.51	279	0.009	0.70469±3	0.70467
M16	secondary minerals 16–19% relict cpx	11.8	443	0.027	0.70482±3	0.70475
M17	Similar to M15, M16	14.5	338	0.043	0.70496±3	0.70485
M18	Less altered basalt; mainly chl	15.4	310	0.050	0.70506±4	0.70494
	pseudomorphs after ol				(0.70519±2)	
M19	Some growth of sphene and albite	11.6	274	0.042	0.70505±5	0.70495
					(0.70514±4)	
M20	Similar to M18	2.62	340	0.008	0.70515±2	0.70513
					(0.70525±6)	
M20	Pyroxene	(0.26)	(51.9)	(0.005	(0.70520±2)	(0.70519)
M21	Similar to M18	9.08	299	0.030	0.70519±3	0.70511
M22	Similar to M15, M16	14.6	311	0.047	0.70531±3	0.70519
		(11.6)	(129)	(0.090)	(0.70538±4)	(0.70516)
M24	Calcite and laumontite vein	2*	169*	0.012	0.70521±6	0.70519
M64	Laumontite vein	2*	359*	0.006	0.70467±3	0.70466
M65	Altered basalt adjacent to M64				0.70505±4	
M75A	Alteration pod: chl-laumontite-cc				0.70421±6	
M75B	Alteration pod: cc, laumontite				0.70427±4	

Rb and Sr contents determined by isotope dilution (±1%) unless postscripted* which denotes XRF analysis. chl, chlorite; plag, plagioclase; qtz, quartz; sph, shene; cpx, clinopyroxene; cc, calcerous; ol, olivine.

† present day $^{87}Sr/^{86}Sr$ ratio ±2 s.e.m. E–A = 0.70807±3 and NBS 987 = 0.71029±3 during this study.

‡ Initial Sr composition calculated assuming an age of 60 Myr and $\lambda = 1.39\times10^{-11}$ yr^{-1} for ^{87}Rb decay.

Results in brackets indicate that they were determined on samples which had been leached in 6M HCp(see text).

more altered material (see Fig. 1). Moreover as no systematic variation in Sr content in the altered rocks was observed (Fig. 1) it seems that Sr isotope exchange took place without net addition or removal of Sr. Finally these results may provide some constraints on our models for the size of water circulation 'cells'. They suggest that the Sr composition of the fluids in this area was not affected significantly by interaction with either 'old' basement (perhaps 0.5 km beneath the present erosion surface) or Tertiary intrusive rocks which crop out only 1.5 km to the south.

Isotope research at Leeds is funded by NERC. We thank the NERC for research fellowships and NERC studentships during this work, and J. Kramers for his review of this article.

C. J. HAWKESWORTH

Department of Earth Sciences, Leeds University, Leeds, UK

M. A. MORRISON

Department of Geology, Imperial College, Prince Consort Road, London SW7, UK

Received 24 August; accepted 3 October 1978.

1. Hart, S. R., Erlank, A. J. & Kable, E. J. D. *Contr. Miner. Petrol.* **44**, 219–230 (1974).
2. O'Nions, R. K., Carter, S. R., Cohen, R. S., Evenson, N. M. & Hamilton, P. J. *Nature* **273**, 435–438 (1978).
3. Forester, R. W. & Taylor, H. P. *Am. J. Sci.* **277**, 136–177 (1977).
4. Forester, R. W. & Taylor, H. P. *Earth planet. Sci. Lett.* **32**, 11–17 (1976).
5. Morrison, M. A. *Earth planet. Sci. Lett.* **39**, 407–416 (1978).
6. Morrison, M. A. thesis Univ. London (1978).
7. Humphris, S. E., Morrison, M. A. & Thompson, R. N. *Chem. Geol.* **23**, 125–137 (1978).
8. Hawkesworth, C. J. *et al. Earth planet. Sci. Lett.* (in the press).
9. O'Nions, R. K. & Pankhurst, R. J. *Earth planet. Sci. Lett.* **31**, 255–261 (1976).
10. Pankhurst, R. J. Walsh, J. N., Beckinsale, R. D. & Skelhorn, R. R. *Earth planet. Sci. Lett.* **38**, 355–363 (1978).
11. Beckinsale, R. D., Pankhurst, R. J., Skelhorn, R. R. & Walsh, J. N. *Contr. Mineral. Petrol.* **66**, 415–428 (1978).

Field measurements of the rheology of lava

KNOWLEDGE of magma rheology is required to understand many igneous and volcanic processes. Viscosities of silicate melts can be estimated from their chemical composition and temperature using empirical methods[1–3]. However, the validity of these estimates has been established only at supraliquidus temperatures where common igneous melts behave as newtonian fluids. Recent experimental evidence from field and laboratory studies[4–7], has demonstrated that, at subliquidus temperatures, some magmas behave as non-newtonian fluids due to the presence of dispersed crystals and gas bubbles and possibly as a consequence of structuring in the silicate melt. We present data here on the rheological properties of small lava flows which were erupted on Mount Etna in 1975. These data were obtained by field measurements using a new field viscometer[8], a shear vane attached to a torque wrench, a conventional penetrometer and various field techniques for estimating rheological properties from the dimensions of lava flows.

Published field measurements of lava rheology have been obtained using various techniques[4,9–12] which, with the exception of one, assume newtonian rheology and also fail to take account of marked thermal (and therefore rheological) gradients which develop at the margins of active flows. The most accurate lava rheology data to be published[4] indicates that the tholeiite in the Makaopuhi lava lake in 1968 was pseudoplastic with yield strengths of 70 and 120 N m^{-2} measured in two experiments and plastic viscosities of 750 and 650 Pa s respectively. The rotating shear vane used for these measurements, however, is suitable only for stable encrusted lava lakes. Because moving lava flows

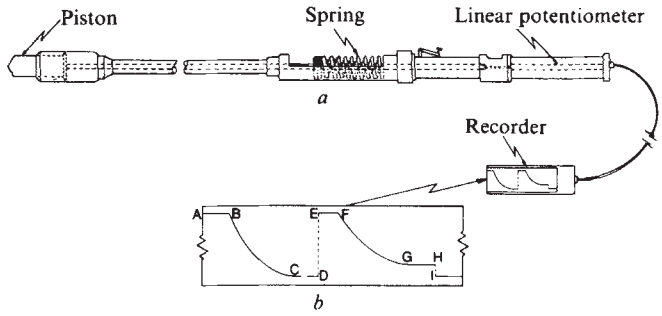


Fig. 1 The Mark 2 viscometer and recorder are shown in (*a*) while an enlarged view of a typical trace from the recorder is shown in (*b*). The Mark 2 is 2 m long and is constructed from high temperature stainless steel. The stainless steel spring used on Etna had a strength of 0.155 N m^{-1}. The recorder, a Cambridge type 72125 single speed hot wire recorder, had a chart speed of 0.0245 m s^{-1}. The traces in (*b*) are characteristic of a newtonian fluid (A–D) and a Bingham material (E–H). For a newtonian fluid, the horizontal distance between A and D is a function of newtonian viscosity. For the Bingham material, GH corresponds to the equilibrium position of the piston in the material; thus the vertical distance HI is related to the dynamic yield strength of the material.

are generally encountered during volcanic eruptions, alternative methods of measuring lava rheology are required. We describe here several techniques with different levels of sophistication which can be used to measure rheological properties of active lava flows.

The Mark 2 field viscometer (Fig. 1) is essentially a penetrometer, but differs from conventional penetrometers in that shearing is confined mainly to the region adjacent to the advancing piston in the interior of the lava. Shearing resistance is not affected by the more viscous cool flow margins as with a conventional penetrometer. In the field the instrument, which is made from BS 321 S20 stainless steel, is preheated by repeated insertions into lava until the temperature of the 'active' end is comparable to that of the lava. It is then inserted to the required depth in the lava and the piston released. A controlled but variable shear rate around the advancing piston is obtained using a compressed spring as the propelling force (Fig. 1) and piston velocity, at a given axial force, is monitored by a linear potentiometer on a hot wire pen recorder. The recorder output is readily converted to shear stress–shear rate data[8]. The yield strength of the lava is obtained directly from the residual force applied by the piston at zero piston velocity (Fig. 1). Thus complete rheological data for the lavas over a wide range of shear rates can be obtained (Fig. 2).

Two additional instruments were used to measure yield strengths on Etna[8]. One is a stainless steel shear vane attached to a torque wrench. After preheating, the vane is inserted into the lava and the torque is increased until the shear vane begins to rotate. Shearing resistance along the shaft is obtained by repeating the measurement at the same depth with the shear vane removed. Thus viscous resistance due to shearing within the lava surrounding the shear vane (and hence the yield strength of the lava) is readily calculated. The third instrument to be used on Etna was a penetrometer. This comprised a 2 m long, 3 cm diameter stainless steel rod, and a pressure transducer which was used to apply a known pressure during insertion. Two other methods were used to acquire rheological data on Etna. These involve (1) measuring the width of newly-formed levees[7,13]; and (2) measuring the dimensions of small lava flows[13,14]. Temperatures were measured during each rheological measurement using a Cr–Al thermocouple.

Rheological data and accompanying temperatures measured on Etna are given in Figs 2 and 3. Figure 2 shows the shear stress–shear rate data for a typical measurement using the Mark 2 viscometer. This measurement was made 3 m downflow from an active bocca at a depth of 0.2 m. The temperature of the lava

0028–0836/78/0276—0383$01.00

© Macmillan Journals Ltd 1978

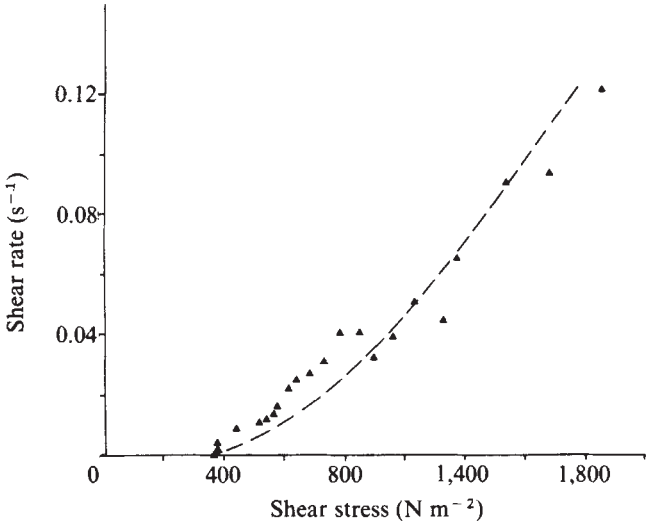


Fig. 2 Rheology of the 1975 Etna lava at 1,086 °C. It can readily be seen that the Etna lava is pseudoplastic with a yield strength of 370(±30) N m^{-2}. If the curve is approximated to a Bingham relationship, the Bingham viscosity is 9,400 ± 1,500 Pa s.

at that point was 1,086 ± 3 °C. The results in Fig. 2 confirm that, at this temperature, the Etna lava had a yield strength of 370 ± 30 N m^{-2}, and it behaved in a pseudoplastic manner. If the curve is approximated to a Bingham relationship, the Bingham or plastic viscosity is 9,400 ± 1,500 Pa s at shear rates $<0.15\ s^{-1}$. Lower plastic viscosities at higher shear rates are possible.

The yield strength of the interior of the 1975 Etna lavas was also calculated using the shear vane, the penetrometer and dimensions of initial levees[8]. Yield strengths measured by these methods are comparable with those calculated using the Mark 2 (Fig. 3), although the consistently higher values obtained by the conventional penetrometer suggest that shear resistance due to the cooler, outer margins of lava flows significantly affected the apparent yield strengths measured using this type of instrument. The relationship between the yield strength of the interior of the 1975 Etna lava and temperature are given by curve 1 in Fig. 3. Four estimates of yield strength were obtained from flow front dimensions. These measurements record the average or 'bulk' rheological properties of the lava which has a low yield strength in the hot interior and a high yield strength in the cool margins. The relationship between the bulk yield strength and the temperature of the isothermal interior of the Etna lavas is shown by curve 2 in Fig. 3. From curves 1 and 2 it seems that the bulk yield strengths of small advancing Etna lava flows are ~2.5 times greater than their hot interiors.

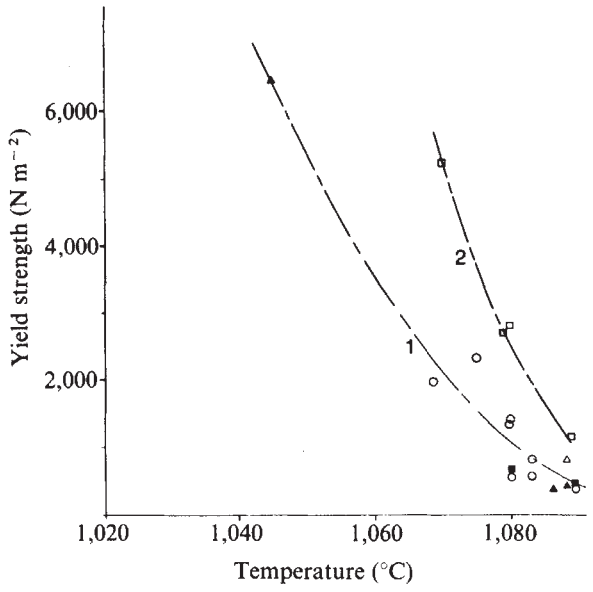


Fig. 3 Relationship between the central temperature and yield strength of specific points along several 1975 Etna lava flows. Curve 1 indicates the trend for the interior of lava flows, while curve 2 indicates the possible relationship for the fronts of advancing flows. It is postulated that curve 1 will tend, asymptotically, to a yield strength of zero at supraliquidus temperatures. ▲, Mark 2; ○, Shear vane; △, penetrometer; □, flow front data; ■, initial levee data.

Etna was the first volcano for which non-newtonian behaviour of lava was proposed[15]. Our results confirm that, on eruption, the 1975 sub-terminal Etna lavas were pseudoplastic with a yield strength. Compared with the tholeiite in Makaopuhi lava lake in 1968[4], the Etna lava flows had higher crystal contents (modal analyses: 35% for Hawaii and 45% for Etna) and lower effusion temperatures (1,135 °C for Hawaii and 1,090 °C for Etna). The higher crystal concentrations and lower effusion temperatures of the Etna lavas are, at least partly, responsible for the higher yield strengths and plastic viscosities of the Etna lava flows[3].

For the volcanologist, non-newtonian behaviour of lava has many implications, both in understanding the flow behaviour of lava flows and in interpreting their morphology. For example, as the apparent viscosity of a pseudoplastic material decreases with increasing shear rate, a fast-moving lava flow will seem to be more fluid than when it moves more slowly. Consequently, if the viscosity of lava flows is calculated from the commonly used Jeffrey's equation[12], Fig. 2 shows that lower apparent viscosities will be calculated for faster-moving lava flows.

For the petrologist, magma rheology is of considerable importance in many processes, such as ascent of magmas, crystal settling[16,17], convection in magmas and xenolith transport[18]. Depending on the reasons for the development of a yield strength in magmas, some, or all, of these processes may be affected by the possible non-newtonian properties of magmas. If the yield strength results entirely from the interaction effects of the phenocryst phase, then only relatively large-scale processes will be affected by the yield strength, for example the settling of xenoliths. On the other hand, if the yield strength results from the formation of a microcrystalline mosaic[19], liquid–liquid immiscibility[20], or structural reorganisation of the melt[21], small-scale processes will also be affected. In particular, crystal settling will not be possible in certain magmas. The reasons for the development of a yield strength in lava flows are investigated more fully elsewhere[22].

H. PINKERTON

Lunar and Planetary Unit,
Department of Environmental Sciences,
University of Lancaster,
Lancaster, UK

R. S. J. SPARKS

Graduate School of Oceanography,
University of Rhode Island,
Kingston, Rhode Island 02881

Received 18 August; accepted 12 October 1978.

1. Bottinga, Y. & Weill, D. F. *Am. J. Sci.* **272**, 438–475 (1972).
2. Shaw, H. R. *Am. J. Sci.* **272**, 870–893 (1972).
3. Scarfe, C. M. *Nature Phys. Sci.* **241**, 101–102 (1973).
4. Shaw, H. R., Wright, T. L., Peck, D. L. & Okamura, R. *Am. J. Sci.* **266**, 225–264 (1968).
5. Shaw, H. R. *J. Petrol.* **10**, 510–535 (1969).
6. Murase, T. & McBirney, A. R. *Geol. Soc. Am. Bull.* **84**, 3563–3592, (1973).
7. Sparks, R. S. J., Pinkerton, H. & Hulme, G. *Geology* **4**, 269–271 (1976).
8. Pinkerton, H. thesis, Univ. Lancaster (1978).
9. Palmer, H. S. *Hawaiian Volcano Observatory Bull.* **15**, 1–4 (1927).
10. Takahasi, R. & Nagata, T. *Tokyo Earthquake Res. Inst. Bull.* **15**, 1047–1053 (1937).
11. Macdonald, G. A. *Geol. Soc. Am. Bull.* **74**, 1071–1078 (1963).
12. Booth, B. & Self, S. *Phil. Trans. R. Soc.* A**274**, 99–106 (1973).
13. Hulme, G. *Geophys. J. R. astr. Soc.* **39**, 361–383 (1974).

14. Johnson, A. M. *Physical Processes in Geology* (Freeman Cooper, California, 1970).
15. Robson, G. R. *Nature* **216**, 251–252 (1967).
16. Shaw, H. R. *Am. J. Sci.* **263**, 120–152 (1965).
17. McBirney, A. R. & Noyes, R. M. *Am. J. Sci.* (in the press).
18. Sparks, R. S. J., Pinkerton, H. & Macdonald, R. *Earth planet. Sci. Lett.* **35**, 234–238 (1977).
19. Taniguchi, H. *Sci. Rep. Tohoku Univ. Ser.* 3 **12**, 189–237 (1974).
20. Philpotts, A. R. *Lithos* **5**, 1–18 (1972).
21. Cornelisse, Ch. W., Visser, T. J. M., Stein, H. N. & Stevels, J. M. *J. Non. Crystall. Solids* **1**, 150–162 (1969).
22. Pinkerton, H. & Sparks, R. S. J. *Nature* **276**, 385–386 (1978).

Effect of degassing on rheology of basaltic lava

BEFORE eruption, most basaltic lava flows lose volatiles, either by Hawaiian—or Strombolian-type fire-fountaining. This gas loss may be of fundamental importance in explaining the observed non-newtonian characteristics of many lava flows. The model proposed here also helps to explain the variations in the surface morphology of many basaltic lava flows.

The loss of most common volatile components increases the viscosity of silicate melts[1,2], and this has been cited as a cause of moderate changes in the viscosity of lava flows[3]. However, an even more important consequence of gas loss is a sudden undercooling of the lava triggering an all-pervasive growth of quench crystals in the lava. This process leads to a rapid increase in viscosity and the development of a high yield strength in the lava.

Degassing of magma leads to undercooling of lava by three processes. First, decompression of the gas phase cools the total magma system adiabatically. For water, ~20 °C cooling occurs per 1 wt% released (ref. 4 and L. Wilson, unpublished calculations). Second, the solution of gases suppresses the liquidus of a melt and alters the order of appearance of mineral phases[5]. Hence, sudden gas loss upsets equilibrium and the melt will generally become undercooled with respect to its anhydrous liquidus. Experimental studies on water solubility and the effect of increasing water pressure on liquidus temperatures[2,5,6] indicate that common types of basalt which contain 1% water would be saturated at 300–400 bar. Such a basalt would be undercooled, by this process, between 30 and 50 °C, if brought rapidly to the surface and then degassed. Carmichael *et al.*[7], on the basis of thermodynamic arguments, also deduced that the crystallisation temperatures of silicate phases would increase on volatile loss. Third, fire-fountaining results in some mixing with the atmosphere and opportunity for radiative heat loss. Some or all of this cooled magma may then be erupted as a lava flow. Swanson and Fabbi[8] have shown that fountain height increases with greater volatile content in Hawaiian fountains which fed lava flows. Consequently, the high fountains associated with large volatile contents would result in increased atmospheric cooling. Data from chilled pillow margins indicate that mid-ocean ridge basalts have water contents ranging from 0.25 to 1.0% with a tendency for alkali basalts to have higher contents than tholeiitic basalts[9–11], whereas basalts from calc-alkaline island arcs and back-arc basins seem to contain even higher water contents[12,13]. Taking a mean water loss from basalts of 1 wt%, we estimate that, as a result of the above three mechanisms, this loss would, typically, cause the lava flow to become undercooled by at least 50 °C. The amount of undercooling will increase with increasing gas loss.

Kirkpatrick[14–16] has shown that crystal nucleation and growth rates both increase with higher values of undercooling of a melt, attain a maximum at several tens of °C, and then decrease. For example, maximum nucleation and growth rates in basalt occur at 50–100 °C undercooling, that is, the same range of undercooling which we predict results from gas loss in many basalts. From Kirkpatrick's studies we estimate that plagioclase crystals will grow at rates of 0.1–1.0 μs^{-1} in a lava undercooled by 50 °C. Using a conservative value of 0.1 μs^{-1}, a 20 μm crystal could grow in 3.5 min. Growth and nucleation will continue until the residual melt has attained equilibrium. With large undercoolings the rapid growth of an extensive mosaic of crystallites can be anticipated in the early stages of flow. It also follows that strongly undercooled lavas (high original gas content) will require more extensive crystallisation to re-establish equilibrium at the extrusion temperature. As basalt lava flows commonly have emplacement times of hours, days or even weeks, gas loss thus results in an all-pervasive growth of quench crystals in the early stages of flow.

This mechanism is, therefore, far more effective in changing lava rheology than atmospheric cooling which penetrates into lavas only slowly because of their poor thermal conductivity[18–23]. In addition, crystallisation changes the composition of the residual melt which, in most cases, would become more siliceous. This effect, together with the sudden appearance of crystal dispersants and the accompanying loss of volatiles, would result in a greatly increased viscosity. More important, the lava would develop a yield strength. An aphyric magma which, before eruption, may have behaved in a newtonian manner will change to non-newtonian behaviour (develop a yield strength) as quench crystallisation takes place. A lava containing many phenocrysts will increase both in yield strength and viscosity.

Several morphological features of lava flows have been interpreted as resulting from non-newtonian rheology[18,24,25]. The formation of levees and channels and the ability of lavas to cease to flow on steep slopes has been attributed to the yield strength of lavas[18,25]. The formation of *aa* surfaces has been interpreted as being due to the inability of surface tension forces to overcome the yield strength[18].

Although lava flows with *aa* surfaces and levee and channel morphology are characteristic of high yield strength lavas, pahoehoe surfaces have also been observed forming on similar lavas, for example, Etna, 1975 (refs. 20, 23). On Etna the main factor controlling the type of lava which formed was effusion rate, pahoehoe surfaces forming on the low effusion rate lava flows. However, this generalisation cannot be extended to all other volcanoes as pahoehoe surfaces can also form on extensive flows with high inferred effusion rates[26]. Apparently therefore, when lava flows have a moderate to high yield strength, pahoehoe lavas form only at low effusion rates, whereas *aa* lavas form when effusion rates exceed a critical value ($2\times10^{-3}\ m^3\ s^{-1}$ on Etna in 1975). On the other hand, when lavas behave as newtonian fluids, or when their yield strength is low, pahoehoe flows form, even at high effusion rates. Although the yield strength of the Etna lavas results, at least partly, from their high phenocryst content, certain flows with a low phenocryst content show features indicative of a high yield strength. Examples are the 1961 Askja basalt lava[27] and the 1875 Sveinagja basalt[28].

These paradoxes in the morphology of basalts can be resolved by the mechanism outlined here. Magmas with high volatile contents will produce strongly undercooled degassed lava. These lavas will develop a high yield strength due to quench crystallisation and will form channelled *aa* flows except at low effusion rate. Magmas with low volatile contents will not be strongly undercooled so will not change rheology substantially on extrusion; they would have either low yield strengths or even behave as newtonian fluids. Formation of pahoehoe flows, even with high discharge rates, would result. Evidence for high volatile content is indicated in the Askja and Sveinagja eruptions directly, by the intense fire-fountaining observed and, indirectly, by the evolved nature of the basalts indicating extensive fractionation of anhydrous phases[28]. It is interesting that the 1961 Askja lavas turned from *aa* to pahoehoe flows later in the eruption, when the intensity of fire-fountaining waned.

The most sluggish basalts are commonly found in island arc settings. These basalts are generally much more viscous than basalts associated with the spreading environments. They invariably produce *aa* or block lava flows and are usually accompanied by violent fire-fountaining, for example, the 1971–72 eruption of Villarica, Chile. Calc-alkaline basalts are

0028–0836/78/0276—0385$01.00
© Macmillan Journals Ltd 1978

generally believed to have high water contents[12,13,29] compared with other common basalt types. Our proposal qualitatively explains these relationships. These basalts are more strongly undercooled after gas loss and consequently have much higher yield strengths and apparent viscosities.

There are several implications of these proposals. First, models of lava flow movement will be required to consider quench crystallisation due to gas loss as a major cause of downflow changes in rheology. Second, the rheology of active lava flows will generally not reflect the rheology of the magma before eruption, and there is a fundamental change in rheology accompanying gas loss. Hence, although many lava flows are non-newtonian this does not necessarily imply that all ascending magmas are non-newtonian. Finally, deductions based on the yield strength of lava from studies of flow morphology cannot be related simply to magma composition[18,30], but must also be related to gas content.

R. S. J. SPARKS

Graduate School of Oceanography,
University of Rhode Island,
Kingston, Rhode Island 02881

H. PINKERTON

Lunar and Planetary Unit
Department of Environmental Sciences,
University of Lancaster, Lancaster, UK

Received 18 August; accepted 2 October 1978.

1. Shaw H. R. *J. geophys. Res.* **68,** 6337–6343 (1963).
2. Scarfe, C. M. *Nature phys. Sci.* **241,** 101 (1973).
3. Macdonald, G. A. *Volcanoes* (Prentice-Hall, New Jersey, 1972).
4. Boyd, F. R. *Geol. Soc. Am. Bull.* **72,** 387–426 (1961).
5. Yoder, H. S. & Tilley, C. E. *J. Petrol.* **3,** 342–532 (1962).
6. Hamilton, D. L., Burnham, C. W. & Osborn, E. F. *J. Petrol.* **5,** 21–39 (1964).
7. Carmichael, I. S. E., Turner, F. J. & Verhoogen, J. *Igneous Petrology* (McGraw-Hill, New York, 1974).
8. Swanson, D. A. & Fabbi, B. P. *J. Res. US geol. Surv.* **1,** 649–658 (1973).
9. Moore, J. G. *Am. J. Sci.* **263,** 40–52 (1965).
10. Moore, J. G. *Contr. Mineral. Petrol.* **28,** 272–279 (1970).
11. Moore, J. G. & Schilling, J. G: *Contr. Mineral. Petrol.* **41,** 105–118 (1973).
12. Meunow, D., Delaney, J. R., Meijer, A. & Liu, N. *EOS Abstr. Trans. Am. geophys. Un.* **58,** 530 (1977).
13. Anderson, A. T. *Bull. Volcanol.* **37,** 530–552 (1973).
14. Kirkpatrick, R. J. *Am. Mineral.* **60,** 798–814 (1975).
15. Kirkpatrick, R. J. *J. geophys. Res.* **81,** 2565–2571 (1976).
16. Kirkpatrick, R. J. *Geol. Soc. Am. Bull.* **88,** 78–84 (1977).
17. Taniguchi, H. *Sci. Rep. Tohoku Univ.* **12,** 189–237 (1974).
18. Hulme, G. *Geophys. J. R. astr. Soc.* **39,** 361–383 (1974).
19. Jaeger, J. C. *Basalts* (ed. Hess, H. H.) **2,** 503–536 (Interscience, New York, 1968).
20. Pinkerton, H. & Sparks, R. S. J. *J. Volcanol. Geothermal Res.* **1,** 167–182 (1976).
21. Peterson, D. W. & Swanson, D. A. *Speleology* **2,** 209 (1974).
22. Shaw, H. R., Wright, T. L., Peck, D. L. & Okamura, R. *Am. J. Sci.* **266,** 225–264, (1968).
23. Pinkerton, H. & Sparks, R. S. J. *Nature* **276,** 383–385 (1978).
24. Sparks, R. S. J. Pinkerton, H. & Hulme, G. *Geology* **4,** 269–271 (1976).
25. Johnson, A. V. *Physical Processes in Geology* (Freeman, Cooper, San Franscisco, 1970).
26. Walker, G. P. L. *Phil. Trans. R. Soc.* A **274,** 107–118 (1973).
27. Sigvaldason, G. E. *Beitr. Miner. Petrogr.* **10,** 263–274 (1964).
28. Sigurdsson, H. & Sparks, R. S. J. *Bull. Volcanol.* (in the press).
29. Yoder, M. S. *Generation of Basaltic Magma* (National Acadamy of Sciences, Washington, 1976).
30. Hulme, G. & Fielder, G. *Phil. Trans. R. Soc.* A**285,** 227–234 (1976).

Pre-ceramic maize on the north-central coast of Peru

AN archaeological site, located near the beach line in a desert part of the Huarmey Valley, 292 km north of Lima, in the department of Ancash, province of Casma, district of Huarmey, 78°10′21″ W, 10°02′45″ S in Peru, was first observed by D. H. Kelley in 1957. It had been previously found by E. P. Lanning who identified it as a pre-ceramic site. It was first designated Huarmey North 1 and it appears as such in previous literature; we are assigning it its real name, Los Gavilanes. We report here ^{14}C dating of maize plant, tassel and cob material from this pre-ceramic site in Huarmey, Peru, at 3,750 ± 100 yr BP.

Charred stones of an underground structure where the associated maize material was found has a date of 4,800 ± 500 yr BP, by thermoluminescence analysis. The corn material was identified as being almost entirely of the Proto-Confite Morocho race, a modern counterpart of which exists today only in the central highlands of Peru. No evidence was found of maize races originating outside of the central Andean region.

A first sounding made by Kelley produced some maize material, which was examined by P. C. Mangelsdorf. Further excavations produced 238 different maize cob, tassel and plant parts. These findings were in a pre-ceramic context[1].

^{14}C measurements made on the maize material gave erratic results, which made the validity of the dating doubtful, so that a preliminary report[2] gave only general information, and no dating. It was found later that corn itself is an inadequate organic material for carbon dating, because growing corn has a most unusual rate of adsorption of ^{14}C and "... dates of maize remains are known to be consistently too young .... This is due to the fact that *Zea mays* together with many other plants (mainly grasses) from semi-arid and arid regions utilise the recently discovered 4-carbon or Hatch–Slack photosynthetic pathway and not the more widespread 3-carbon or Calvin pathway for carbon dioxide fixation. Such C-4 species are not common in temperate regions; the syndrome generally constitutes an adaptation to a hot and dry environment" (ref. 3). It was also found that the technique of handling and treating the maize material presumably had not been appropriate to eliminate contamination by masses of crystallised carbonates which were later found adhering to the surface of many corn cobs at the site, and which could have obscured the radio-carbon results.

In 1974, with support of the Royal Ontario Museum of Canada, a new excavation was made, which yielded the results presented here and which permitted us to re-evaluate earlier information[2] and prepare a new report[4]. The site is located on the right of the Huarmey river banks, bordering on a fossil lagoon, which was formed between barriers of old beach lines over an old stream bed. Alyssian winds cause constant movement of shifting sands that have covered the site completely. On the surface only stone accumulations are visible, some of which could be circular structures, arranged in a confusing pattern.

The 1974 excavation was made in one of the circular structures. This was only partially completed because of the large volumes of sand that had to be removed. The excavation yielded an irregular inverted truncated cone, the walls of which are built with irregular stones, 9 m in diameter at the top, and 1.30 m deep. On the walls of this structure abundant residues of maize leaves were found. A cut on one of the sides of the structures made in order to establish its nature, showed abundant refuse mixed with the stone.

Preliminary analysis has also shown the existence of human and animal bones, predominating among the latter are a shore and sea fauna, with scarce representation of *Camelidae*, *Cervidae* and *Canidae*. There is also abundance of molluscs, primarily *Bivalvia*, *Gasteropoda* and *Amphineura*. Plant remains of *Distichlis spicata* L., *Schirpus* sp., *Tillandsia* sp., *Inga* sp., *Pachyrizus* sp., *Schinus molle* L., *Gossypium barbadense* L., *Cucurbita* sp., *Lagenaria siceraria* (Mol.) Standl, and *Ambrosia* sp. were also found.

The artefacts found at the site are mostly lithic. Tentatively, artefacts can be distinguished as made from cobblestone: these are basically strikers, choppers and chopping tools. Circular flat pebbles, worked along the edges by unifacial percussion and whose function is still unknown, are also characteristic and distinctive. Tools obtained from flakes are scrapers, notched pieces and some cores were used like scrapers; there is also a sizeable proportion of flakes used without further preparation. There are also textile fragments made by the technique of twining, typical of the pre-ceramic period. Some coprolites were also found and these are being studied at present.

On the basis of the charcoal pieces found associated with the debris, a ^{14}C dating was obtained of 3,750 ± 110 yr BP (GIF 3564, Laboratoire du Radiocarbone du Commisariat à l'Ener-

gie Atomique CNRS). Charred stones of the excavated structure were dated by thermoluminescence analysis at 4,800 ± 500 yr BP (BOR 20, Laboratoire de Cristallographie et Physique Cristalline, University of Bordeux).

All evidence points to the Los Gavilanes site as belonging to the final Peruvian pre-ceramic period, that described by Lanning[5] as Pre-ceramic 6, and in which corn apparently played an important part.

Abundant maize cob, stalk and tassel material was recovered from the 1974 excavation made at the Los Gavilanes site. Nine intact corn cobs and 12 cob fragments were examined and measured (Fig. 1). They ranged in length from 2.5 to 9.4 cm, with diameters ranging from 0.7 to 1.4 cm. About 70% of the ears and cob fragments have 8 rows of kernels, with all the remainder having ten rows except one with 12 rows. The cupules of the cobs are distributed rather laxly, are rather deep, and are 2.5 mm in width and 4.5 to 6.0 mm in length, some with prominent cupule wings, pilosity of the cupules mostly low to medium, but a few with high pilosity. The kernels may have been isodiametric, about 3 mm in size, of the popcorn type, with 12 to 13 kernels in each row. The glumes are rather long, the external ones being semi-hard, coriaceous and the internal ones very soft and papery; they are either a quarter or half tunicate. Their colour is brownish red, and a few are purple (probable allele arrangement A_1A_2BP1).

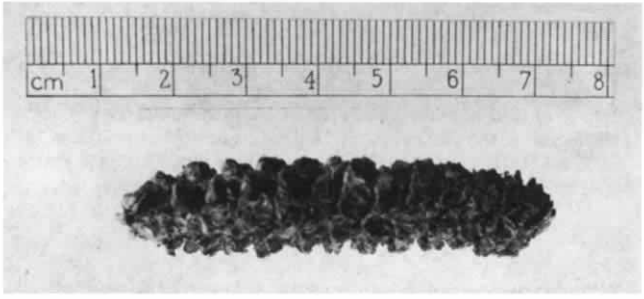


Fig. 1 Corn cob from Los Gavilanes site.

Husk material has a venation index of 7 to 9, the veins being very prominent. This is a rather high index[6], representative of primitive maize material. One of the husks exhibits a very definite purple colour. This colour type is prevalent today in the highland area of Peru, and has been shown[7] to have a definite selective advantage at high altitudes. Its presence in high frequencies in the Los Gavilanes population is a good indication of possible transfer of ancestral seeds from the highlands to the coast of Peru.

The cob peduncle is rather slender as in most highland popcorns. Stalk diameter is small, between 0.6–0.8 cm, and about 0.85 cm including leaf sheath.

Tassel fragments (Fig. 2) indicate a very lax type, such as found today in both primitive and advanced highland maize races of Peru. Condensation index of the tassels[6] ranges from 0.59 to 1.2, with the majority of the tassels falling within the shorter range of 0.58–0.8.

Unquestionably there is a single predominating maize race in this complex, an early form of Confite Morocho, designated as Proto-Confite Morocho[6]. A few cobs exhibit a fair degree of introgression of another race Proto-Confite Chavinense, better shown in one ear with some fasciation, and the only one in the cache to have 12 rows.

The abundance of plant and tassel material at the site would seem to indicate that the complete plants were harvested in the cultivated part of the valley and brought to the site tied in bundles, possibly carried on people's backs.

It is possible that the structure that has been opened was used as a silo. An interesting observation in this connection is that all

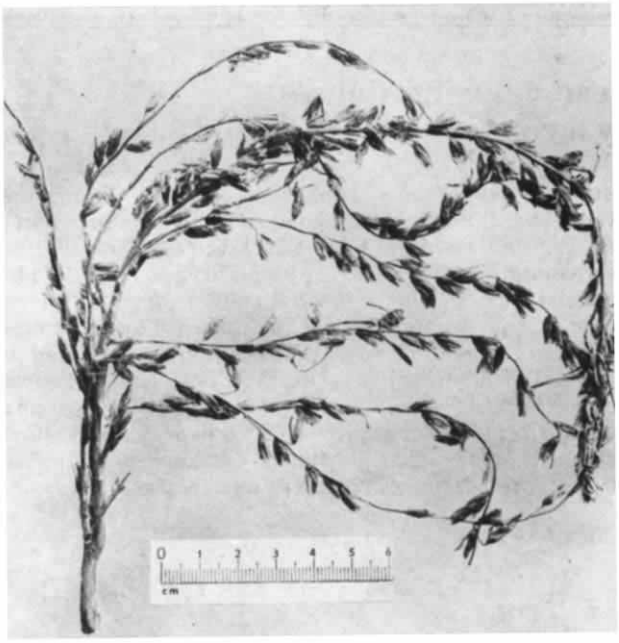


Fig. 2 Complete tassel from Los Gavilanes site.

the stones of the sides of the excavated structure have been burned. Similar storage pits lined with stones were found by A.G. full of ears of corn in sandy hills near Lima. The corn ears were protected from insects and rodents by filling the spaces between them in the pits with sand. Similar modern storage systems using sand were still in use in 1950 in the Huarmey Valley, and were examined at that time by A.G.

The implication of the find is that maize was cultivated on the coast of Peru in the pre-ceramic period, about 4,000 years ago. This maize is undoubtedly related to the Confite Morocho race presently existing in the highlands of Peru and one of the most primitive races in existence. This find tends to confirm the postulation of both the origin of primitive coastal maize of Peru, as migrating from the Peruvian highlands[6], and the origin of cultivation and domestication of maize in the Central Andean region as an independent process from the domestication of maize in Meso-America[8]. There is no evidence on this site at this early period of introduction of maize resembling the primitive types that were being cultivated in Mexico contemporaneously.

We thank Virginia Popper and Elizabeth S. Wing for their help with identifications, Mario Peña for help with malacological data, Max Schvoerer for thermoluminescence dating, Georgette Delibrias for ^{14}C dating and Claude Chauchat who took part in ground research.

ALEXANDER GROBMAN

Centro Internacional de Agricultura Tropical, Apartado Aéreo 6713, Cali, Colombia

DUCCIO BONAVIA

Universidad Peruana Cayetano Heredia, Departamento de Biología, Apartado 2563, Lima, Perú

Received 10 July; accepted 22 September 1978.

1. Kelley, D. & Bonavia, D. *Ñawpa Pacha* **1**, 39–41 (1963).
2. Mangelsdorf, P. C. & Cámara-Hernández, J. *Maize Genetics Cooperation Newsletter* **41**, 47–48 (1967).
3. Vogel, J. C. & Van Der Merwe, N. J. *American Antiquity* **42**: 2, 238–242 (1977).
4. Grobman, A., Bonavia, D., Kelley, D., Mangelsdorf, P. C. & Cámara-Hernández, J. *Botanical Museum Leaflets*, **25**: 8, 221–242 (1977).
5. Lanning, E. P. *Peru before the Incas* (Prentice-Hall, Englewood Cliffs, New Jersey, 1967).
6. Grobman, A., Salhuana, W., Sevilla, R. & Mangelsdorf, P. C. *Races of Maize in Peru* (Nat. Acad. Sci., Publ. No. 915, Washington, 1961).
7. Greenblat, I. M. *Maize Genetics Cooperation Newsletter* **42**, 144–145 (1968).
8. Bonavia, D. & Grobman, A. El Origen del Maíz Andino (Fastchrift: Herman Trimborn. Instituto Anthropos de San Agustín, Bonn, West Germany, in the press).

Lead detoxification by a copper-tolerant isopod

THE acquisition of metal tolerance by freshwater and marine invertebrates in environments polluted by mine waste in south-west England has now been established[1,2]. Generally, for invertebrates such tolerance has been shown to be specific for particular metals[1], although the literature suggests that co-tolerance may operate for copper and silver in the polychaete *Nereis diversicolor*[3] and for zinc and nickel in the grass *Agrostis tenuis*[4,5]. We show here that co-tolerance operates in the isopod crustacean *Asellus meridianus*, in which tolerance to copper seems to confer tolerance to lead. Two populations of *A. meridianus* from differing habitats are also shown to use different methods for achieving lead tolerance.

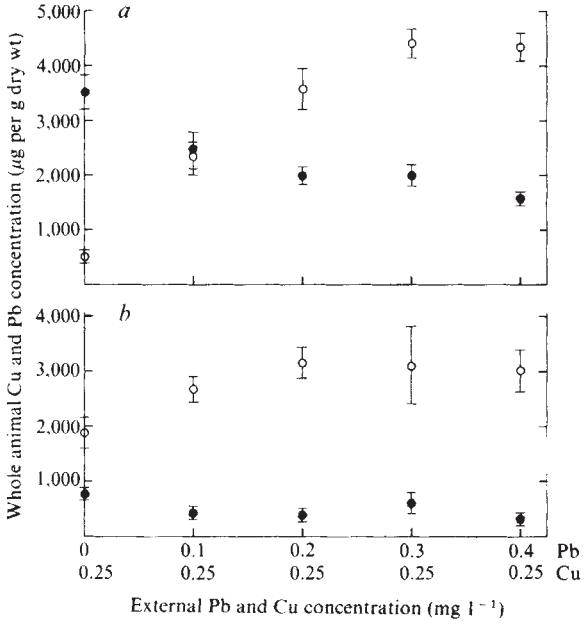


Fig. 1 The effect of increasing concentrations of lead (0.1–0.4 mg l^{-1}) and a constant copper concentration (0.25 mg l^{-1}) on uptake of copper (●) and lead (○) by *A. meridianus* from *a*, the Hayle and *b*, the Gannel, after 14 d exposure to test solutions at 15 °C ± 1 °C. Experimental techniques followed those described in detail elsewhere[2]. Test solutions were made up using Analar grade $CuSO_4$ and $Pb(NO_3)_2$ in a freshwater medium; solutions were renewed every 48 h throughout the experiment. Twenty animals were used at each dilution and batches of five whole animals were wet digested in Aristar HNO_3 and analysed for copper and lead with an Evans electroselenium 240 atomic absorption spectrophotometer. Each point is the mean of four batch samples; bars represent the s.e.m.

In the present study, two populations of *A. meridianus* Rac were studied from two rivers, the Hayle and the Gannel, which, respectively, drain disused copper and lead mines in Cornwall, U.K. Previous work[2] had established copper and lead tolerance in animals from the River Hayle, where 'total' copper concentrations in the water range between 0.10 and 0.25 mg l^{-1} and levels of lead are low, <0.01 mg l^{-1}. A marked lead tolerance had also been observed in animals from the River Gannel, where total lead concentrations in the water range between 0.19 and 0.35 mg l^{-1}. No copper tolerance was observed in animals from the latter site, where total copper levels in the water are 0.01–0.04 mg l^{-1}.

Animals from both populations were exposed to solutions containing copper and lead alone and to mixed solutions containing the two metals in various proportions. The uptake of copper by Hayle animals is considerably reduced by the addition of increasing levels of lead, which are accumulated readily (Fig. 1*a*). Uptake of copper by the Gannel population in similar conditions is lower than that of Hayle animals and is apparently unaffected by the addition of increasing concentrations of lead. Lead uptake by these animals is also lower than that shown by the Hayle population (Fig. 1*b*).

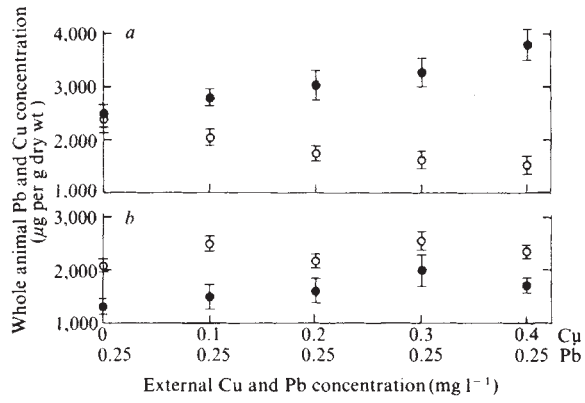


Fig. 2 The effect of increasing concentrations of copper (0.1–0.4 mg l^{-1}) and a constant lead concentration (0.25 mg l^{-1}) on uptake of lead (○) and copper (●) by *A. meridianus* from *a*, the Hayle and *b*, the Gannel, after 14 d exposure to test solutions at 15 °C ± 1 °C. Each point is the mean of four batch samples; bars represent the s.e.m.

When experimental conditions are reversed and animals are exposed to a constant lead concentration and increasing copper levels, uptake of lead by Hayle animals is reduced at higher copper concentrations (Fig. 2*a*). Reduction of the amount of metal accumulated by these animals is not as pronounced as in the previous experiment (Fig. 1*a*). Similarly, uptake of copper is not as marked as uptake of lead in the first experiment, suggesting that lead may be more preferentially bound than copper. Lead uptake by the Gannel population is apparently unaffected by increasing copper concentrations and copper uptake is low compared with that of Hayle animals (Fig. 2*b*).

Earlier work[2,5,6] had established that the hepatopancreas served as a principal storage organ for trace metals in isopod crustaceans. Ultrastructural and subsequent X-ray microanalytical studies of the hepatopancreas of *A. meridianus* from the Hayle exposed to lead suggest that both copper and lead are present at the same sites within the cells of the hepatopancreas (Fig. 3*a*–*c*). It is apparent that both copper and lead compete for sites in the hepatopancreas of this tolerant isopod, as both chemical analysis (Fig. 4) and microanalysis confirm a relationship between the two metals. X-ray microanalysis further demonstrates that lead is present only in the 'cuprosomes' of exposed animals. It is suggested that lead is more readily bound than copper, which is released into the lumen of the hepatopancreas, probably as a result of apocrine secretion (Fig. 3*a*) in which the apical end of the cell is lost together with the granules[8], in response to an incoming dose of lead. Similar preferential binding of lead over copper has also been observed in earthworm chloragosomes (Ireland, personal communication).

The ability of copper-tolerant isopods from the Hayle to 'detoxify' lead by storing the metal in 'cuprosomes' at the expense of copper indicates that metal tolerance in invertebrates may not be as specific as had originally been supposed. Also, the mechanisms operating for particular metals, in this case lead and

© Macmillan Journals Ltd 1978

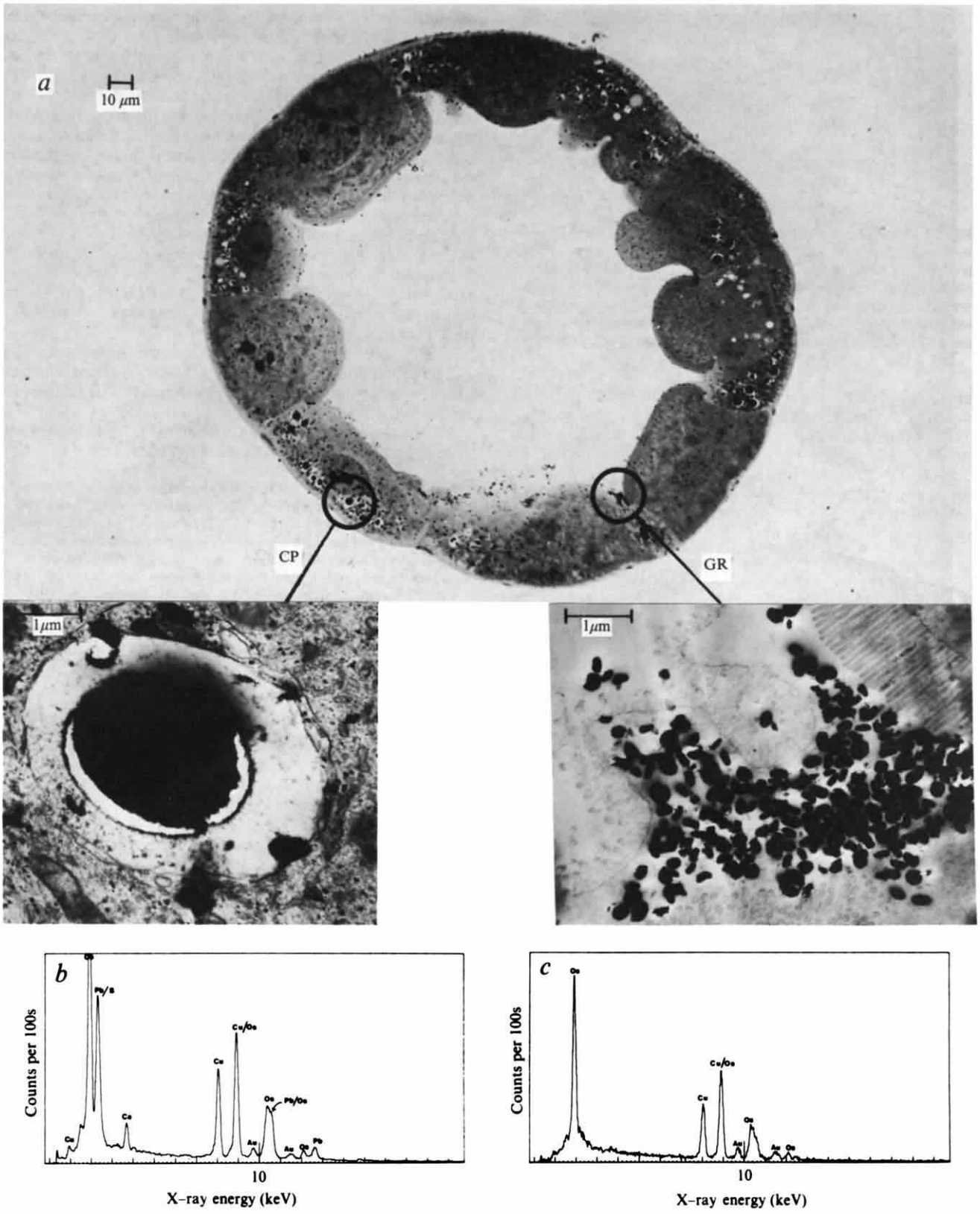


Fig. 3 Uptake of lead by isopods from the River Hayle exposed to 0.5 mg l^{-1} Pb for 10 d at 15 °C ± 1 °C. *a*, Photomicrograph showing 1-μm section of the hepatopancreas, stained with toluidine blue, of an exposed Hayle animal. Insets show electronmicrographs of densely staining 'cuprosomes' (CP) around the border of the cells and granular material (GR) being released into the lumen of the hepatopancreas. The hepatopancreatic caeca of lead-exposed animals were fixed in gluteraldehyde and osmium tetroxide, dehydrated and embedded in Araldite. 1-μm sections were cut and stained in toluidine blue and thin sections (600–1,000 Å) were then mounted on copper grids for ultrastructural analysis. The sections were viewed in an EM6-b (AEI Scientific Apparatus) electron microscope. *b*, The elemental spectrum of a cuprosome (CP) of a lead-exposed animal, showing peaks for Os (fixative), Au (grid) and constituents (Pb, Cu, Ca, S) of cuprosome. *c*, The elemental spectrum of granular material in the lumen of the hepatopancreas of a lead-exposed animal showing peaks for Os (fixative), Au (grid) and constituents of granular material (Cu). Hepatopancreatic caeca were fixed as above and sections (600–1,000 Å) mounted on gold grids for analysis by the electron microscope micro-analyser EMMA 4 (AEI Scientific Apparatus). Full vertical scale is 18,300 counts; in both cases the displayed spectrum takes into account the spectrum for the surrounding cytoplasm, which has been subtracted from the original.

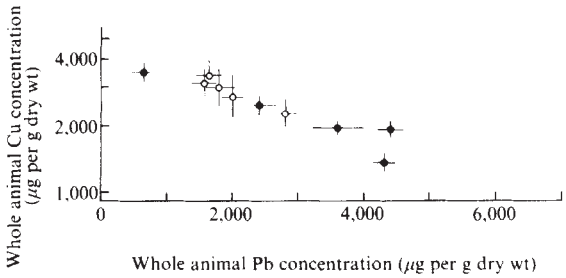


Fig. 4 Levels of copper plotted against levels of lead accumulated by whole animals from the Hayle during exposure to copper and lead alone and to mixed solutions of copper and lead. ●, Points obtained during exposure of animals to a constant Cu concentration and increasing Pb concentrations; ○, points obtained during exposure of animals to a constant Pb concentration and increasing Cu concentrations. Each point is the mean of four batch samples; bars represent the s.e.m.

copper, in different populations of the same species are not necessarily similar, as indicated by the ability of Gannel animals to restrict the uptake of both copper and lead into the hepatopancreas.

I thank Dr T. Appleton for assistance with X-ray microanalysis and R. Hewitt for assistance in preparation of material for electron microscopy. The work was supported by the Natural Environment Research Council.

B. E. BROWN

The Dove Marine Laboratory,
Cullercoats, Tyne and Wear, UK

Received 7 August; accepted 16 October 1978.

1. Bryan, G. W. & Hummerstone, L. G. *J. mar. biol. Ass. U.K.* **51**, 845–863 (1971); **53**, 839–857 (1973).
2. Brown, B. E. *Water Res.* **10**, 555–559 (1976); *Freshwat. Biol.* **7**, 235–244 (1977).
3. Bryan, G. W. in *Effects of Pollutants on Aquatic Organisms* (ed. Lockwood, A. P. M.) 7–34 (Cambridge University Press, 1976).
4. Turner, R. G. in *Ecological Aspects of Mineral Nutrition of Plants*, (ed. Rorison, I. H.); *Br. ecol. Soc. Symp.* **9**, 399–410 (Blackwell, Oxford, 1969).
5. Gregory, R. P. G. thesis, Univ. Wales (1964).
6. Wieser, W. *Am. Zool.* **8**, 495–506 (1968).
7. Wieser, W. & Klima, J. *Mikroskopie* **24**, 1–9 (1969).
8. Humbert, W. *Z. Morph. Ökol. Tiere* **78**, 93–109 (1974).

Transient tritanopia experiment in blue cone monochromacy

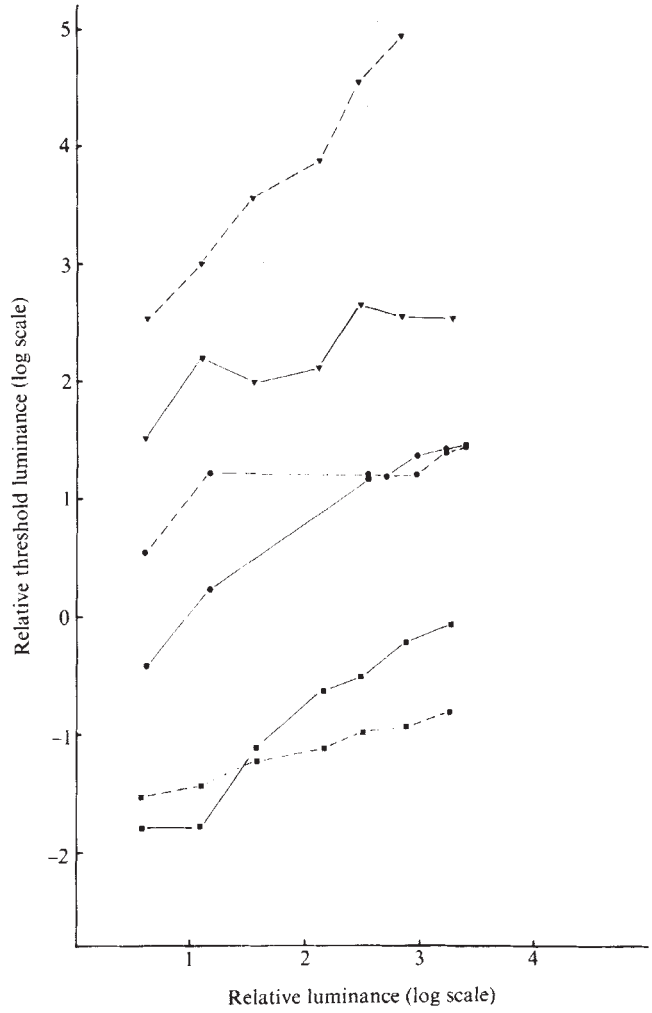
WHEN the eye has been adapted to a bright yellow light, a marked loss of sensitivity to short-wavelength stimuli may be recorded immediately after the extinction of the adapting field. This phenomenon, first described by Stiles[1], was termed transient tritanopia by Mollon and Polden[2]. They supposed that a recovering long-wavelength mechanism may inhibit or otherwise suppress the blue cone signal during early dark adaptation. Transient tritanopia was found in one protanope and, in a modified form, in one deuteranope. This indicates that each of the red or green mechanisms or both in combination may give rise to the inhibition effect. It is of particular interest in this connection to know if this kind of inhibition is present in persons without red or green cones. In blue cone monochromacy, a unique kind of colour vision anomaly, there are only blue cones and no functioning red or green cones. We report here the experimental results obtained with such a person where no suppression effect upon the blue mechanism could be demonstrated. The results are compared with those of a typical rod monochromat and a person with normal vision.

The blue cone monochromat (BTW), a 17-yr-old male, had visual acuity of 6/60 in each eye. He had pendular nystagmus and moderate photophobia. The presence of functioning blue cones was ascertained during the clinical examination and by spectral sensitivity studies[3]. By means of two modified Prado Universal projectors, a monochromatic circular test spot of 450 nm and 1° angular size was presented on a 30° yellow adapting field projected on a $BaSO_4$ screen. The viewing distance was 0.40 m. The test spot was presented about 1° above a fixation mark. The field of adaptation was produced by a Wratten 12 gelatine filter. Every 18 s the adapting field was extinguished for 3 s and a single test flash, lasting 30 ms, was presented 400 ms after the extinction of the field. Stimulus duration was controlled by electromechanical shutters. Maxwellian-view presentation was not preferred because of the subject's nystagmus. Preliminary experiments did not show significant differences either using artificial pupils or with maximally dilated pupils. The experiments were therefore performed with natural pupils. During the short period of 400 ms after extinction of the adapting field the pupil is not likely to show any significant change[4].

At photopic adaptation levels, the normal observer shows a distinct loss of threshold sensitivity in the dark interval. The blue cone monochromat subject, however, showed no significant loss of sensitivity at any level of adaptation (Fig. 1).

A slight increase of threshold sensitivity at lower adaptation levels is seen in the normal observer and is due to rod intrusion

Fig. 1 The transient tritanopia experiment in three kinds of observers: a normal trichromat (■), a blue cone monochromat (●) and a rod monochromat (▼). Ordinate, relative threshold luminance at which the blue test spot is detected; abscissa, relative luminance of adapting field. - - - -, log threshold luminance recorded on a steady yellow field. ——, log threshold luminance 400 ms after the adapting field has been turned off. Luminance is given relative to 1 cd m^{-2}. The curves for the blue cone monochromat and the rod monochromat have been displaced 2 and 4 log units respectively along the ordinate axis.



© Macmillan Journals Ltd 1978

consequent to the parafoveal stimulation. This increase of sensitivity is more pronounced in the blue cone monochromat, probably because his nystagmus causes a greater participation of rod receptors.

As rod activity thus influences the response functions, it is of interest to study the results obtained in the same conditions in a rod monochromat, that is, a person without ordinary cones. The results of the rod monochromat are shown in Fig. 1 by the top curves. At low adaptation levels, the response pattern is similar to that of the blue cone monochromat, but it differs significantly at higher adaptation levels. The ratio between threshold values in the light and dark intervals increases with the luminance of the background light. At high luminances the threshold sensitivity regains more than 2 log units during the short period of 400 ms after extinction of the background light. This fast recovery rate is essentially different from the responses of the two other observers having both cones and rods in their retinas.

When acting without the influence of other receptors, the rod mechanism thus has a very great ability of recovering after exposure to the yellow light. By the presence of blue cones this ability of fast recovering is completely lost, as the threshold values are now at the same level irrespective of the background light being on or off. When green and red cones are present, there is also a suppression of the blue mechanism giving further loss of threshold sensitivity in the dark interval. This, which is the transient tritanopia, is demonstrated by the normal observer.

The lack of transient tritanopia in the blue cone monochromat supports the theory of Mollon and Polden that this phenomenon in normal persons is an effect due to the inhibition of the blue cone mechanism by the other cones.

E. HANSEN

Department of Ophthalmology,
University Hospital,
Rikshospitalet,
Oslo 1

T. SEIM
B. T. OLSEN

Biophysics Department,
Institute of Physics,
University of Oslo,
P.O. Box 1048, Blindern,
Oslo 3, Norway

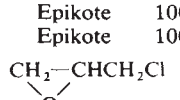
Received 7 June; accepted 29 September 1978.

1. Stiles, W. S. *Documenta Ophthal.* **3**, 138–169 (1949); in *Mechanisms of Colour Vision* (Academic, New York, 1978).
2. Mollon, J. D. & Polden, P. G. *Nature*, **258**, 421–422 (1975); *Nature* **259**, 570–572 (1976); *Phil. Trans. R. Soc.* **278**, 207–240 (1977).
3. Hansen, E. *Acta ophthal.* (in the press).
4. Le Grand, Y. in *Light, Colour and Vision* 98 (Wiley, New York, 1957).

Mutagenic action of aromatic epoxy resins

EPOXY RESINS are used extensively for protective coatings and in paints and adhesives; the consumption of epoxy resins during 1973 was 190,000 metric tons, which is an almost twofold increase over a period of 10 years[1]. About 90% of the epoxy resins used are manufactured by condensation of two molecules of epichlorhydrin with one or more molecules of bisphenol A (2,2-bis (4-hydroxyphenyl)-propane). The mutagenic action of simple epoxides such as ethylenoxide and epichlorhydrin was first shown in fruit flies[2] and has subsequently been demonstrated in a wide range of organisms. Epoxy resins are diepoxides and bifunctional alkylating agents, and bifunctional alkylating epoxides are known to be mutagenic[3,4]. Mutagens are

Table 1 Compounds tested by the *Salmonella*/microsome mutagenicity test

		MW	*n*
Epikote 828	(E 828)	370	0
Epikote 1001	(E 1001)	950	1
Epikote 1004	(E 1004)	1,800	3
			
Epichlorhydrin	(ECH)	92.5	

The Epikote resins were supplied by Shell Chemicals. According to that firm, Epikote 828 is basically the diglycidylether of bisphenol A, as indicated by the formula, plus a small proportion of higher molecular weight material. The higher molecular weight resins contain some resin of lower molecular weight. Epichlorhydrin was from Fluka.

suspected of having a carcinogenic effect because of the relationship between carcinogenicity and mutagenicity[5] and most diepoxides are carcinogenic in mice and rats[6]. We report here that aromatic epoxy resins are mutagenic in *Salmonella typhimurium*, and may thus represent a cancer risk in man.

Three aromatic epoxy resins were tested by the plate incorporation assay with histidine-requiring mutants of *S. typhimurium*. The test was carried out as described by Ames *et al.*[7]. Epichlorhydrin (ECH) was included as positive control (see Table 1). The epoxy resins are all mutagenic in tester strain TA 100 (Fig. 1).

To ensure that mutagenic action was not due to interaction between the resins and dimethyl sulphoxide (DMSO), which is normally used as solvent in the plate incorporation assay, we also tested an aqueous emulsion made by ultrasonic emulsification of E 828. The aqueous emulsion was mutagenic to *S.typhimurium* TA 100, but to a lesser extent than in assays where DMSO was used as solvent.

E 828 was the most active compound of the resins tested, and both E 828 and E 1001 were more mutagenic than the positive control, ECH. The solubility of the resins decreases as the molecular weight increases. The resins were suspended in the agar medium in the amounts shown in Fig. 1. However, as the resins were not in solution, the amounts given may not express the true doses.

Comparison of the first column of Table 2 with Fig. 1 shows that the mutagenic action of ECH in tester strain TA 1535 was about the same as that demonstrated with TA 100. In contrast, the mutagenic activity of the resins was very low compared with that observed using TA 100. This indicates that the mutagenic action of the resins is not caused solely by the possible content of unreacted ECH in the resins.

Assays were carried out to determine the mutagenic effect of epoxy resin metabolites. After addition to the plates of rat liver microsomes, prepared from rat liver activated with sodium phenobarbital and containing NADP, the mutagenic activity of the epoxy resins and ECH was reduced almost 10-fold in assays with TA 100.

When the same assays were carried out with TA 1535, E 828 was clearly activated by rat liver enzymes in the presence of NADP (compare the second and first columns in Table 2). In contrast, this resin was inactivated in assays where NADP was omitted (see third column in Table 2). The results obtained with E 1001 showed the same tendency.

As in the assays with TA 100, the mutagenic activity of ECH in assays with TA 1535 was reduced almost 10-fold by rat liver enzymes (see Table 2). This reduction in activity was independent of the presence of NADP. The inactivation of epoxy resins and ECH by microsomes without NADP was enzymatic, since the addition of heat-inactivated microsomes to the plates gave the same results as shown in the first column of Table 2.

0028–0836/78/0276—0391$01.00

© Macmillan Journals Ltd 1978

It is known from studies of polyaromatic hydrocarbons and styrene oxide that the microsomal epoxide hydrase catalyses the hydration of epoxides to diols, and that the soluble fraction glutathione-*S*-epoxide transferase catalyses the conjugation of epoxides with glutathione[8]. These two enzymes are active in rat liver preparations without NADP. The resulting metabolites are considered to be non-mutagenic or less mutagenic than the epoxide[9,10].

The activation of the aromatic epoxy resins as found in TA 1535 is not caused by decomposition of the compounds and liberation of bisphenol A. Bisphenol A was tested for mutagenicity, both with and without the addition of liver microsomes and was not active in tester strains TA 1535 or TA 100.

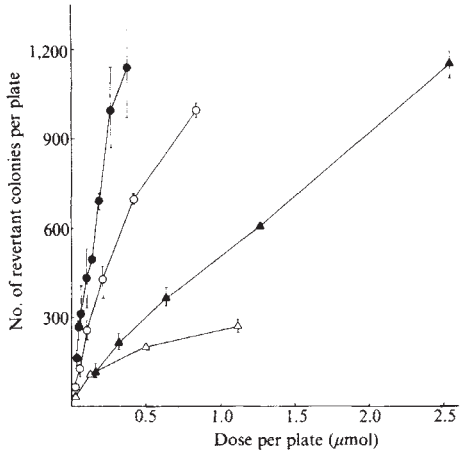


Fig. 1 Mutagenicity of aromatic epoxy resins and ECH for *S. typhimurium* TA 100 (revert by base-pair substitution). The epoxides were dissolved in 100 μl dimethyl sulphoxide (Merck, spectrometric grade). Bacteria and dissolved test agent were incorporated into the agar overlay. No S-9 mix was added. The number of spontaneous revertant colonies was estimated at 250. Each point is an average of numbers scored on three plates. The limits of the vertical lines indicate the lowest and highest values obtained with each dose. ●, E 828; ○, E 1001; △, E 1004; ▲, ECH (positive control).

Extensive studies of benzo(*a*) pyrene metabolites show that the combined epoxide hydrase–monooxygenase system leads to the formation of highly mutagenic diol-epoxide[11]. The monooxygenase system is dependent on NADP. Metabolites of this type might explain why NADP-dependent enzymes play a role in the formation of metabolites which are more mutagenic than the original epoxy resins to tester strain TA 1535, as shown in Table 2.

The demonstration of a mutagenic effect of aromatic epoxy resins in *S. typhimurium* indicates a genetic hazard, including a cancer risk, for humans exposed to these compounds. The fact that epoxy resins can be metabolised to mutagenic active compounds enhances the genetic risk from aromatic epoxy resins. Man is exposed to the resins during their manufacture and use either by skin contact with the compound or by inhalation of air contaminated with droplets or powder particles of epoxy resin. It has been demonstrated that topical aromatic epoxy resins (MW 350) are toxic to rats[12]. Therefore, it seems that the resins can penetrate the skin and be absorbed by animals.

Aromatic epoxy resins have been tested for carcinogenicity. After one skin exposure to 900 μM aromatic epoxy resin (MW 395), one skin tumour and three malignant lymphomas appeared in 50 mice[13]. In a study with only 40 mice, no tumours were found after skin painting with undiluted resin[14] but results are inadequate since none of the animals was alive after 24 months. In a study by Hine *et al.*[15], four malignant sarcoma tumours were produced at the site of injection after administration to 30 rats.

Table 2 Activity of aromatic epoxy resin metabolites ascertained by the *Salmonella* microsome mutagenicity test

Compound	Dose (μmol per plate)	−S-9 mix	+S-9 mix	+S-9 mix (−NADP)
E 828	0.54	ND	39	ND
	1.35	ND	109	ND
	5.40	58	521	ND
	13.5	79	996	20
	54.0	162	ND	14
E 1001	2.1	24	39	9
ECH	0.254	216	ND	ND
	0.635	474	ND	ND
	2.54	1,418	34	35
	6.35	ND	152	105
	25.4	ND	911	731

The figures represent the numbers of revertant colonies per plate produced by various doses of aromatic expoxy resins and epichlorhydrin (ECH) in tester strain TA 1535 (reverting by base-pair substitution). The spontaneous revertents (36) are subtracted. Each figure is an average of numbers scored on 3–10 plates in two independent assays. ND, not determined. The plate incorporation assay described by Ames *et al*[7]. was followed without modification. The assay was performed with and without the addition of rat liver microsome preparations. To some plates, preparations containing NADP were added, that is S-9 mix described by Ames *et al.*[7]. In other plates S-9 mix without NADP was used. 150 μl S-9 was added to each plate.

Since malignant tumours were demonstrated, and as the number of animals was small and the period of observation short in the above-mentioned studies, there is no basis for concluding, as did Hine *et al.* and others, that aromatic epoxy resins are safe from the point of view of carcinogenicity.

Owing to the long latent period for genetic damage and most cancers in humans[16], it is not possible to identify the genetic or cancerogenic hazards of aromatic epoxy resins by epidemiological studies in the immediate future, since consumption on any scale may not have taken place until the 1960s.

To prevent cancer and genetic damage in humans it is necessary to minimise the exposure to substances such as aromatic epoxy resins which have been shown to be mutagenic in bacteria, and thus must be considered as potential mutagens and carcinogens in human beings.

This work was supported by the Danish MRC.

MARGRETHE ANDERSEN*
PAULI KIEL
HENRY LARSEN
JETTE MAXILD

Department of Microbiology,
The Royal Danish School of Pharmacy,
2 Universitetsparken,
DK-2100 Copenhagen Ø, Denmark

Received 15 May; accepted 19 September 1978.

*Names in alphabetical order.

1. Sorin, B. in *Health Hazards from Epoxy* (eds Ehrner-Samuel, H. & Sjöholm, O.) 6–10 (Arbetarskyddsstyrelsen, Stockholm, 1974).
2. Rapoport, I. A. *Dokl. Akad. Nauk SSSR* **60**, 469–472 (1948).
3. Sram, R. J., Cerna, M. & Kucerova, M. *Biol. Zbl.* **95**, 451–462 (1976).
4. McCann, J., Choi, E., Yamasaki, E. & Ames, B. N. *Proc. natn. Acad. Sci. U.S.A.* **72**, 5135–5139 (1975).
5. McCann, J. & Ames, B. N. *Proc. natn. Acad. Sci. U.S.A.* **73**, 950–954 (1976).
6. Van Duuren, B. L. *Ann. N.Y. Acad. Sci.* **163**, 633–651 (1969).
7. Ames, B. N., McCann, J. & Yamasaki, E. *Mutat. Res.* **31**, 347–364 (1975).
8. James, M. O., Fouts, J. R. & Bend, J. R. *Biochem. Pharmac.* **25**, 187–193 (1976).
9. Milvy, P. & Garro, A. J. *Mutat. Res.* **40**, 15–18 (1976).
10. Bresnich, E. in *In Vitro Metabolic Activation in Mutagenesis Testing* (eds de Serres, F. J., Fouts, J. R., Bend, J. R. & Philpot, R. M.) 91–105 (North-Holland, Amsterdam, 1976).
11. Wood, A. in *In Vitro Metabolic Activation in Mutagenesis* (eds de Serres, F. J., Fouts, J. R., Bend, J. R. & Philpot, R. M.) 179–197 (North-Holland, Amsterdam, 1976).
12. Cornish, H. H. & Block, W. D. A.M.A. *Archs ind. Hlth* **20**, 44/390–52/398 (1959).
13. Kotin, P. & Falk, H. L. *Radiat. Res. Suppl.* **3**, 193–211 (1963).
14. Weil, C. S., Condra, N., Haun C. & Striegel, J. A. *Am. ind. Hyg. Ass. J.* **24**, 305–325 (1963).
15. Hine, C. H., Guzman, R. J., Coursey, M. M., Wellington, J. S. & Anderson, H. H. *Cancer Res.* **18**, 20–26 (1958).
16. Higginson, J. *Proc. Eur. Soc. Toxic.* **17**, 104–112 (1976).

T cell origin of human macromolecular insoluble cold globulin

THYMUS-DERIVED (T) lymphocytes are a heterogeneous cell population in terms of their function and the surface antigens they express. In the mouse, a subpopulation of sensitised T cells is cytotoxic for allogeneic or modified syngeneic target cells by virtue of recognition of histocompatability antigens, that is, H-2K and H-2D (refs 1 and 2). These cytotoxic T lymphocytes express a specific antigen of the Ly antigen series, Ly-2,3 (refs 3 and 4). Another subpopulation of T cells mediate responses in the mixed lymphocyte reaction and the graft-versus-host reaction[4,5]. These cells recognise immune response associated antigen (Ia) and possess another antigen of the Ly series, Ly-1 (refs 4 and 5). In addition, separate classes of T cells mediate helper and suppressor regulatory effects on antibody response

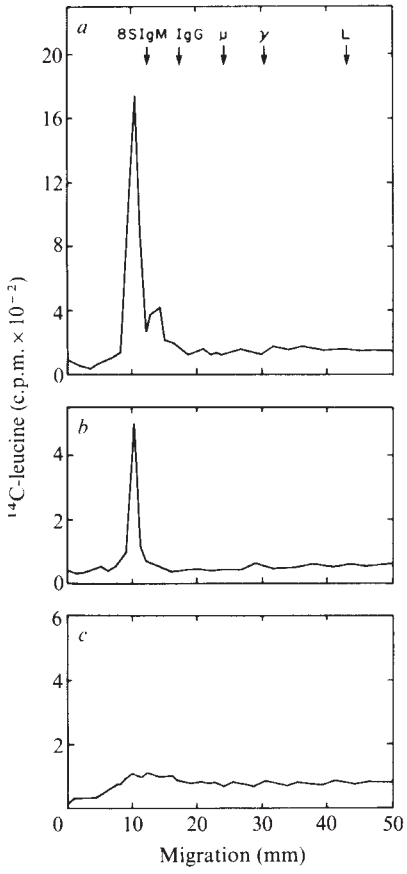


Fig. 1 Cold precipitation of ^{14}C-leucine labelled human MICG protein from lysates of: *a*, PBL (36×10^6); *b*, T cells (12×10^6); *c*, B cells (12×10^6). Precipitates were reduced and subjected to SDS–PAGE (5%). Top of gel is on the left side of figure. Arrows indicate the position of ^{14}C-leucine labelled markers (derived from mouse plasmacytoma cells of the appropriate class) which were electrophoresed on companion gels[13]. 8S IgM unreduced monomer IgM (MW 200,000); IgG, unreduced intact IgG; μ, IgM heavy chains; γ, IgG heavy chains; L, light chains.

and also have distinctive Ly membrane markers[4,6,7]. Although membrane proteins such as immunoglobulin and B-cell antigens are specific to bone marrow-derived (B) lymphocytes in humans, there are few well characterised proteins that serve to distinguish human T cells[8–12]. We have recently identified a glycoprotein synthesised in mouse thymocytes and splenocytes[13,14]. This protein has been termed macromolecular insoluble cold globulin (MICG) to describe its major physicochemical properties. It has a molecular weight of 225,000, is insoluble in the cold in non-ionic detergents and has the electrophoretic mobility of a

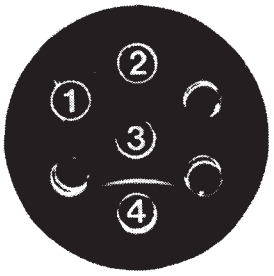
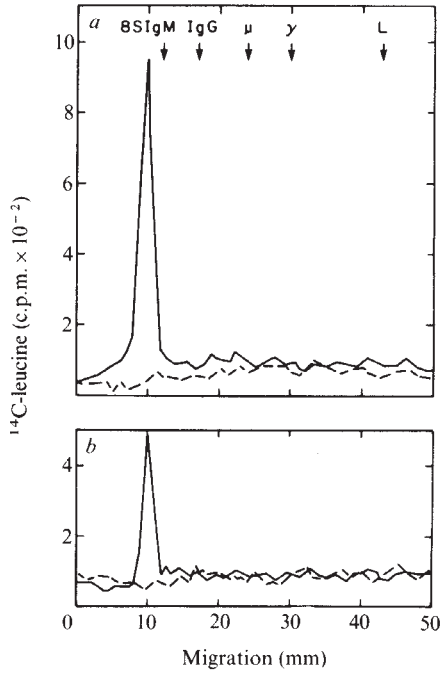


Fig. 2 Ouchterlony analysis (double immunodiffusion). Wells: 1, PBL lysate in 0.5% Triton X-100 and 2 M urea; 2, rabbit antiserum to human MICG; 3, rabbit antiserum to human cold insoluble globulin (given by M. Mosesson); 4, normal human plasma.

β globulin. MICG has also been shown to be the exclusive product of T lymphocytes[15]. This report describes the characterisation of a similar protein in human T cells.

Peripheral blood lymphocytes (PBL) were separated from heparinised whole blood by Ficoll-Hypaque centrifugation (FHC)[16]. T lymphocytes were isolated from PBL by rosette formation with sheep red blood cells (SRBC) and sedimentation on FHC[17]. The lymphocytes from the SRBC pellet (after lysis of SRBC) were used as a source of T cells. They contained less than 2% surface immunoglobulin-positive cells and 97% of these cells formed rosettes with SRBC. B lymphocytes were isolated from the B cell-enriched fraction (unrosetted cells) at the Ficoll–medium interphase by application to Sephadex G-200 anti-human Fab columns[18,19]. Bound cells, eluted by competitive inhibition with human IgG, were 98% surface immunoglobulin-positive and less than 1% formed rosettes with SRBC. Surface immunoglobulin-positive cells were determined by fluorescence microscopy using rabbit anti-human light chain antiserum labelled with fluorescein isothiocyanate[20].

Fig. 3 Antiserum precipitation of radiolabelled lysates from PBL, T cells and B cells. Lymphocytes, labelled with ^{14}C-leucine, were lysed in isotonic buffer[13], *p*H 7.2 and 0.5% (v/v) with respect to desoxycholate, for 20 min at 4 °C. Lysates were immunoprecipitated using antiserum (250 μl) in antibody excess for 16 h at 4 °C. Samples were reduced and treated on SDS–PAGE (5%). *a*, PBL (75×10^6) precipitated with either anti-MICG antiserum (——) or normal rabbit serum (- - - -); *b*, T cells (40×10^6) (——) or B cells (40×10^6) (- - - -) precipitated with anti-MICG antiserum.



0028–0836/78/0276—0393$01.00
© Macmillan Journals Ltd 1978

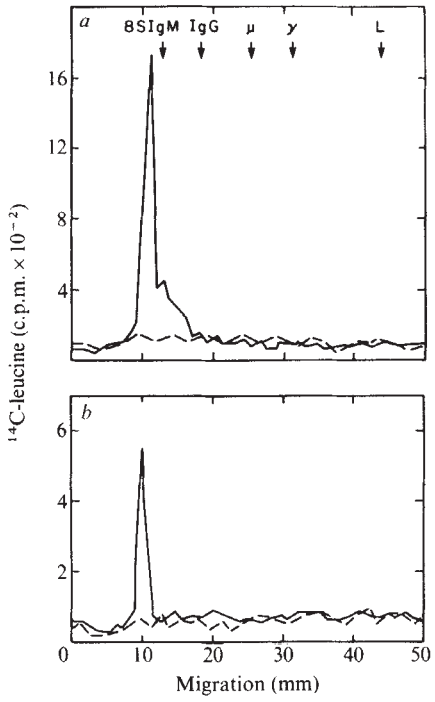


Fig. 4 Specificity of antihuman MICG antiserum. Radiolabelled (^{14}C-leucine) PBL lysates were precipitated and electrophoresed on 5% SDS gels after reduction: *a*, PBL (50×10^6) were lysed in NP-40 and cold precipitated (——) at 4 °C for 2 d and the supernatant reacted with anti-MICG antiserum (250 μl) in antibody excess (- - - -). *b*, Lysates from PBL (30×10^6), prepared as in Fig. 3, were reacted with either unadsorbed anti-MICG antiserum (——) or anti-MICG adsorbed with T cells (300×10^6 T cells per ml antiserum) for 1 h at 37 °C (- - - -).

Lymphocytes, radiolabelled with 10 μCi of ^{14}C-leucine, were lysed in Nonidet P-40 (NP-40) and the lysates cold precipitated for 2 d at 4 °C in detergent buffer[14]. The washed precipitates were dissolved in 2% SDS–8 M urea–phosphate buffer for 1 h at 37 °C, reduced with 0.5 M 2-mercaptoethanol and alkylated with 0.65 M iodoacetamide. The samples were then electrophoresed on 5% SDS–polyacrylamide gels, and the gels sliced and counted[14]. Cold precipitated PBL lysates demonstrated a major peak of radioactivity with a molecular weight of 225,000 (Fig. 1*a*), compared with appropriate standards[14]. Incubation of cell lysates for 2 d at 22 °C did not induce the precipitation of this macromolecule. Similar to mouse MICG, the 225,000 MW protein had the electrophoretic mobility of a β globulin.

To investigate the cell of origin of this protein, PBL were separated into T and B lymphocyte subpopulations and radiolabelled lysates analysed for cold insolubility. As shown in Fig. 1*b* and *c*, only isolated T cells were able to synthesise the cold precipitable macromolecule. There was no radioactivity cold precipitated from B lymphocytes.

The relationship between MICG and human serum proteins was analysed with antiserum to isolated MICG[14]. This antiserum did not react with isolated, radioiodinated proteins, including albumin, α_2 macroglobulin, transferrin, IgG, IgA or IgM, nor did it react with concentrated human plasma or serum in Ouchterlony analysis[14]. In contrast, antiserum to MICG showed a single precipitan band with PBL lysates (Fig. 2). Note also that a variety of monospecific antisera to human serum proteins did not react with PBL lysates in Ouchterlony analysis. These included antisera to α_2 macroglobulin, albumin, transferrin, IgM, IgA, IgG, IgD, IgE, fibrinogen, cold insoluble globulin and whole plasma (Fig. 2).

Monospecificity of this antiserum was also shown by direct precipitation of radiolabelled lysates derived from PBL, T cells and B cells. As MICG is insoluble in the cold in non-ionic detergents, it was necessary to lyse lymphocytes with a detergent in which MICG was soluble. When radiolabelled lymphocytes were lysed in Desoxycholate, the macromolecule remained in solution. After precipitation of these lysates with antiserum to MICG, a single peak of radioactivity was demonstrated on SDS gels for both PBL and T cell lysates (Fig. 3*a* and *b*). The molecular weight of this protein was 225,000, when compared with appropriate standards. In contrast, no radioactivity was precipitated from B cells with anti-MICG antiserum (Fig. 3*b*), nor did normal rabbit serum react with the radiolabelled lysates (Fig. 3*a*).

To determine whether the 225,000 cryoprotein visualised using SDS–polyacrylamide gel electrophoresis (PAGE) was similar to the protein precipitated by anti-MICG antiserum, radiolabelled PBL lysates were cold precipitated and the supernatant allowed to react with antiserum to MICG. Following cold precipitation there was no radioactivity precipitated from the supernatant with anti-MICG antiserum (Fig. 4*a*). The reverse of this experiment, whereby radiolabelled lysates were first precipitated with anti-MICG antiserum and the supernatant cold precipitated, also demonstrated antigenic similarity (not shown).

As a final proof of the specificity of this antiserum for T cells, the antiserum was adsorbed with T cells and reacted with radiolabelled PBL lysates. In these conditions, the antiserum was not reactive with the radiolabelled lysate (Fig. 4*b*). When the antiserum was adsorbed with B lymphocytes, there was no loss of activity against PBL.

The present experiments demonstrate a unique protein present in human T cells. It has similar chemical characteristics to mouse MICG and has been termed human T-MICG. Further studies are in progress to elucidate the biological function of this protein. Certainly, T-MICG serves to distinguish human T cells from B cells.

STEPHEN P. HAUPTMAN
EMIN KANSU

Cardeza Foundation for Hematologic Research,
Department of Medicine,
Thomas Jefferson University,
Philadelphia, Pennsylvania 19107

Received 2 August; accepted 4 October, 1978.

1. Bach, F. H., Widmer, M. B., Bach, M. L. & Klein, J. *J. exp. Med.* **136,** 1430–1444 (1972).
2. Schmitt-Verhulst, A., Sacks, D. H. & Shearer, G. M. *J. exp. Med.* **143,** 211–217 (1976).
3. Cantor, H. & Boyse, E. A. *J. exp. Med.* **141,** 1376–1389 (1975).
4. Cantor, H. & Boyse, E. A. *Immun. Rev.* **33,** 105–145 (1977).
5. Meo, T. *et al. Transplant. Proc.* **7,** (suppl. 1) 127–129 (1975).
6. Mitchison, N. A. *Eur. J. Immun.* **1,** 68–75 (1971).
7. Tse, H. & Dutton, R. W. *J. exp. Med.* **143,** 1199–1210 (1976).
8. Warner, N. L. *Adv. Immun.* **19,** 67–216 (1974).
9. Fu, S. M., Winchester, R. J. & Kunkel, H. J. *J. exp. Med.* **139,** 451–456 (1974).
10. Klareskog, L. *et al. Nature* **265,** 248–251 (1977).
11. Owen, F. L. & Fanger, M. W. *Immunochemistry* **13,** 121–127 (1976).
12. Touraine, J. L., Hadden, J. W. & Good, R. A. *Proc. natn. Acad. Sci. U.S.A.* **74,** 3414–3418 (1977).
13. Hauptman, S. P. & Sobczak, G. *Immun. Commun.* **7,** 69–79 (1978).
14. Hauptman, S. P. *Immunochemistry* (in the press).
15. Hauptman, S. P., Sobczak, G. & Gutterman, I. A. *Immunochemistry* (in the press).
16. Boyum, A. *Scand. J. clin. Lab. Invest.* **21,** (suppl. 97) 77–89 (1968).
17. Greaves, M. F. & Brown, G. *J. Immun.* **112,** 420–423 (1974).
18. Chess, L., MacDermott, R. P. & Schlossman, S. F. *J. Immun.* **113,** 1113–1121 (1974).
19. Rocklin, R. E., MacDermott, R. P., Chess, L., Schlossman, S. F. & David, J. R. *J. exp. Med.* **140,** 1303–1316 (1974).
20. Higmans, W. H., Schuit, H. R. E. & Klein, F. *Clin. exp. Immun.* **4,** 457–472 (1969).

Mode of coupling between hormone receptors and adenylate cyclase elucidated by modulation of membrane fluidity

THE hormone- and neurotransmitter-dependent adenylate cyclases represent transmembrane regulatory systems in which the receptor unit faces the outside of the cell and the cyclase catalytic unit faces the inside. When the hormone or neurotransmitter binds to the receptor, the catalytic unit of the cyclase is activated and produces cyclic AMP at the inner surface of the

0028–0836/78/0276—0394$01.00

© Macmillan Journals Ltd 1978

membrane[1]. It has recently become clear that some hormone receptors are permanently coupled to adenylate cyclase whereas others are not. We demonstrate here that the effects of modulating membrane fluidity on the rate of adenlyate cyclase activation can be used to distinguish between the two modes of coupling.

Activation of adenylate cyclase requires the simultaneous binding of hormone to receptor and intracellular GTP to the guanyl nucleotide regulatory subunit[2]. Furthermore, in the case of the β-adrenergic receptor the hormone signal is terminated concomitantly with the hydrolysis of GTP to GDP and inorganic phosphate at the regulatory site[3]. Thus for the β-adrenergic receptor, and perhaps generally, the continued presence of both hormone and GTP is essential to maintain the steady-state level of active adenylate cyclase[2,3]. This mechanism may also account for the observation that non-hydrolysable GTP analogues induce a permanently active state of the cyclase in the presence of hormone. When the latter occupies the GTP regulatory site, the GTPase step is blocked and the enzyme remains in its inactivated state[2,4,5].

Recent evidence suggests two alternative modes of interaction between the receptor regulatory unit and the catalytic cyclase unit. Either the receptor and the enzyme are mobile and float in the membrane[6-10] (in this case the activation of the catalytic moiety takes place during the encounter between receptor and enzyme) or the receptor and the enzyme are permanently coupled to each other, as a typical soluble regulatory enzyme composed of regulatory and catalytic subunits such as aspartate transcarbanylase (in this case the subunit interactions are similar to those of classic regulatory proteins in solutions). The first mechanism is described in the following scheme:

$$\begin{array}{ccccc} \mathrm{HR+E} & \xrightarrow{k_1} & (\mathrm{HRE}) & \xrightarrow{k_2} & \mathrm{HR+E'} \\ \uparrow\downarrow K_H & & & & \\ \mathrm{H+R} & & & & \end{array} \qquad (1)$$

where H is the hormone, E the enzyme, R the receptor, k_1 the bimolecular rate constant, E′ the activated enzyme, K_H the receptor–hormone dissociation constant and k_2 the first order activation constant. (HRE) indicates the transient ternary complex, the hormone–receptor–enzyme. If the ternary complex is short-lived ($k_2 \gg k_1$), the encounter between hormone-bound receptor and catalytic unit becomes rate limiting. Such a coupling mechanism has been termed 'collision coupling'. The second mechanism (permanent coupling) can be described by the following scheme:

$$\begin{array}{ccc} \mathrm{HRE} & \xrightarrow{k} & \mathrm{HRE'} \\ \uparrow\downarrow K_H & & \\ \mathrm{H+RE} & & \end{array} \qquad (2)$$

where RE is the receptor–enzyme complex, H the hormone, k, the first-order constant of activation and K_H the dissociation constant of hormone to receptor–enzyme complex.

In these two schemes we have eliminated the deactivation process of the active enzyme species. This is achieved in hormone-dependent adenylate cyclases when nonhydrolisable GTP analogues such as guanylylimidodiphosphate (GppNHp) replace GTP in the adenylate cyclase reaction[2,5]. Therefore, in the presence of hormone and GppNHp, the enzyme becomes permanently active and accumulates according to[9,10]:

$$B = A\{1-\exp(-k_{obs}t)\} \qquad (3)$$

In the 'collision coupling' model (equation (1)), A = total concentration of catalytic units and B concentration of the units activated at time t. k_{obs} Is the pseudo first-order rate constant given by:

$$k_{obs} = k_1[\mathrm{R_T}]\frac{[\mathrm{H}]}{K_H+[\mathrm{H}]} \qquad (4)$$

where $[\mathrm{R_T}]$ is the total receptor concentration, k_1 the bimolecular rate constant and the term $[\mathrm{H}]/(K_H+[\mathrm{H}])$ expresses the fraction of hormone-occupied receptors.

The maximal number of the activated catalytic units in that case is independent of receptor concentration, as every receptor can activate an unlimited number of catalytic units through the collision process. It is apparent from equation (4) that k_{obs} is directly proportional to the receptor concentration.

For equation (2), A in equation (3) is the total concentration of receptor–enzyme and B the concentration of the activated receptor–enzyme unit at time t. K_{obs} In this case is given only by:

$$k_{obs} = \frac{k[\mathrm{H}]}{K_H+[\mathrm{H}]} \qquad (5)$$

where k is the first-order rate constant of activation and the term $[\mathrm{H}]/(K_H+[\mathrm{H}])$ is the fractional occupancy of the receptor by the hormone. At saturating hormone concentrations this term is equal to one.

We have recently shown[9,10] that the coupling of turkey erythrocyte adenylate cyclase to the β-adrenoceptor is 'collision coupling' whereas the adenosine receptor is permanently coupled to the enzyme in the same cell[11,12]. This was based on the observation that progessive destruction of the β-adrenergic receptor by a specific affinity label results in a proportional decrease in the rate constant of cyclase activation by the receptor, without reduction in the total number of catalytic units which can be activated. In contrast, progressive destruction of the adenosine receptor by a specific affinity label results in a proportional decrease in the number of catalytic units which can be activated without a change in the rate constant of activation[12]. As the collision coupling mechanism, in contrast to the precoupled mechanism, necessitates a relative movement of the receptor with respect to the catalytic units in the membrane matrix, a different dependence of enzyme activation by the hormone through membrane fluidity is expected for epinephrine compared with adenosine.

To test this hypothesis we modified the fluidity of turkey erythrocyte membranes and measured the kinetics of enzyme activation by epinephrine and by adenosine. Turkey erythrocyte membranes were fluidised progressively by the insertion of *cis*-vaccenic acid which is one of the most efficient fluidising agents[13]. The apparent membrane microviscosity was measured using fluorescence polarisation[14], using 1,6-diphenyl(1,3,5)hexatriene (DPH) as the fluorescence probe. In parallel, we measured the time course of adenylate cyclase activation to its permanently active state, by either *l*-adrenaline or adenosine, in the presence of saturating concentrations of GppNHp. The first-order rate constants for the first-order process of adenylate cyclase activation by the two hormones were derived from equation (3) according to the procedure described elsewhere[9,10].

Figure 1 shows that the rate constant of the epinephrine-dependent activation increases linearly with membrane fluidity, whereas the rate constant of the adenosine-dependent activation is independent of membrane fluidity. This supports the kinetic experiments described above. Furthermore, the linear depencence of k_{obs} on $1/\eta$ in the case of the adrenaline-dependent activation, strongly suggests that the interaction between receptor and enzyme is diffusion controlled.

The inverse relationship between the rate of a diffusion-controlled bimolecular reaction and the viscosity of the medium is well established[15,16]. When the bimolecular reaction occurs in two dimensions the relationship is[17]:

$$k_1 = \frac{k_{obs}}{[\mathrm{R_T}]} = \frac{N_A \cdot 4\pi \cdot D_{R,E}}{\ln\left[\left(\frac{\pi}{4}[\mathrm{R_T}]N_A\right)^{1/2} \cdot \frac{1}{a}\right]} \qquad (6)$$

where k_1 is the bimolecular rate constant (equation (1)), k_{obs} the measured pseudo first-order rate of enzyme activation (equations (3) and (4)), $D_{R,E}$ relative diffusion coefficient, $[\mathrm{R_T}]$ total receptor concentration, N_A Avogadro's number and a the

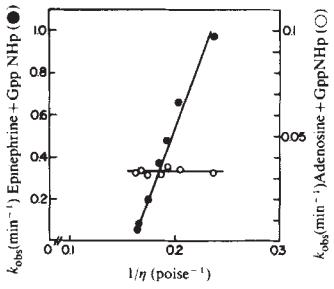


Fig. 1 The dependence of the rate constant of adenylate cyclase activation by *l*-adrenaline and by adenosine on membrane fluidity. Adenosine- and *l*-adrenaline-induced first-order processes of adenylate cyclase activation were measured as a function of increasing membrane fluidity. The kinetics of enzyme activation were determined and analysed as described in the text, using equation (3) to fit the hormone-induced activation curves. The fluidity of the membrane was increased progressively by treating the membrane with *cis*-vaccenic acid as described earlier[13]. The fluidity of the membranes was measured by fluorescence polarisation using DPH as a fluorescence probe[14]. (●), Dependence of k_{obs} for adrenaline activation on membrane fluidity; (○), dependence of k_{obs} for adenosine activation on membrane fluidity.

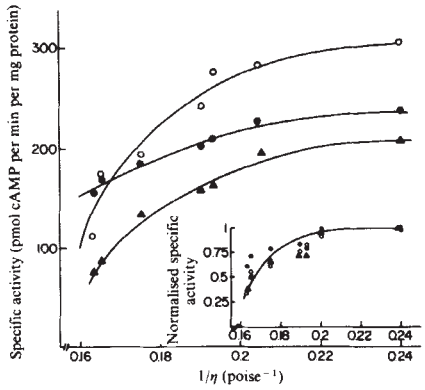


Fig. 2 The dependence of adenylate cyclase specific activity on membrane fluidity. The specific activity of adenylate cyclase as a function of membrane fluidity was measured in the presence of either 0.10 mM *l*-adrenaline plus 0.1 mM GppNHp (○), 0.1 mM adenosine plus 1.0×10^{-4} GppNHp (▲), or in the presence of 10 mM NaF (●). The assay mixture also contained 50 mM Tris-HCl, 6 mM $MgCl_2$, and 1 mM EDTA, *p*H 7.4, as described earlier[29]. Inset: The maximal specific activity for each mode of activation was taken as 1.0 and the relative increase in specific activity as a function of membrane fluidity was plotted. In this way one can compare the dependence of the three modes of activation with each other.

length of receptor to enzyme interaction domain. As the diffusion coefficient is inversely proportional the viscosity of the medium (Stokes'–Einstein relation) one can rewrite equation (6) as:

$$k_{obs}=\frac{N_A\cdot 4\pi[R_T]}{\ln\left[\left(\frac{\pi}{4}[R_T]N_A\right)^{1/2}\cdot\frac{1}{a}\right]}\times\frac{C}{\eta} \qquad (7)$$

where C is a constant and η the viscosity of the medium. From the slope of the plot of k_{obs} against $1/\eta$ the value of C can be calculated. Knowing η one can calculate the value of $D_{R,E}$.

Using the relationship between the bimolecular rate constant k_1 and the viscosity of the medium in a diffusion-controlled reaction in two dimensional space[17], one can calculate that the diffusion coefficient describing the relative motion of the receptor and the enzyme is in the range of 4.0×10^{-11} to 9×10^{-10} cm^2 s^{-1} at 25 °C. The value of k_1 can be obtained from k_{obs} ($k_1\sim1\times10^{13}$ min^{-1} (mol cm^{-2})$^{-1}$ (equation (4)) as $[R_T]$ is known for turkey erythrocytes[18,19]. This range of values for the diffusion coefficient is typical for mobile membrane receptors[20,21]. These values range from 2×10^{-11} to 4×10^{-9} cm^2 s^{-1}. The independence of the adenosine-induced rate of enzyme activation on membrane fluidity (Fig. 1) confirms previous biochemical studies[11,12] which showed that this receptor is permanently coupled to the cyclase.

The validity of the fluorescence depolarisation technique as a measure for membrane viscosity has been questioned[22]. It seems that even if the method does not yield the absolute value of the membrane viscosity these values are directly proportional to the viscosity. The fact that the data presented in this letter confirms the biochemical studies supports the view that the fluorescence polarisation technique is a powerful tool for measuring membrane viscosity. Furthermore, viscosity data which are derived from the rotation of small lipid-soluble molecules in the plane of the membrane approximate viscosity measured in other ways, for example by following diffusion of membrane proteins[23].

Membrane fluidity also affects the catalytic unit directly, as the maximal activity of the enzyme increases as a function of membrane fluidity. From Fig. 2 it is apparent that the adenosine-dependent activity, the adrenaline-dependent activity and the NaF-stimulated activity increase as a function of membrane fluidity in a similar way. Similar effects of membrane viscosity on the adenylate cyclase activity were recently reported[24,25]. The specific activity of other membrane bound systems, such as Ca^{2+}-ATPase[26], (Na^+, K^+)ATPase[27] and the β-galactoside transport system[28] were also found to depend on the membrane fluidity.

The dependence of the rate of adenylate cyclase activation by a hormone receptor on membrane fluidity can be adopted as a general diagnostic test for investigating coupling between a hormone receptor and adenylate cyclase. This approach can be extended to other membrane processes which involve the interaction between a receptor and a catalytic moiety, other than cyclase, such as an ionophore.

This work was supported by the US–Israel Binational Research Foundation.

GILAD RIMON
EMANUAL HANSKI
SERGEI BRAUN
ALEXANDER LEVITZKI

Department of Biological Chemistry,
The Institute of Life Sciences,
The Hebrew University of Jerusalem,
Jerusalem, Israel

Received 5 June; accepted 29 September 1978.

1. Perkins, J. P. *Adv. Cyclic Nucleotide Res.* **3,** 1–64 (1973).
2. Levitzki, A. *Biochem. biophys. Res. Commun.* **74,** 1154–1159 (1977).
3. Cassel, D. & Salinger, Z. *Biochim. biophys. Acta* **452,** 538–551 (1976).
4. Schramm, M. & Rodbell, M. *J. biol. Chem.* **250,** 2232–2237 (1975).
5. Sevilla, N., Steer, M. L. & Levitzki, A. *Biochemistry* **15,** 3493–3499 (1976).
6. Boeynames, J. M. & Dumont, J. E. *J. Cyclic Nucleotide Res.* **1,** 123–142 (1975).
7. Jacobs, S. & Cuatrecasas, P. *Biochim. biophys. Acta* **433,** 482–495 (1976).
8. Levitzki, A. *Biochem. Pharmac.* (in the press).
9. Tolkovsky, A. M. & Levitzki, A. in *Hormones and Cell Regulation*, Vol. 2 (eds Dumont, J. & Nunez, J.) 89–105 (North-Holland, Amsterdam, 1978).
10. Tolkovsky, A. M. & Levitzki, A. *Biochemistry* **17,** 3795–3810 (1978).
11. Tolkovsky, A. M. & Levitzki, A. *Biochemistry* **17,** 3811–3817 (1978).
12. Braun, S. & Levitzki, A. *Biochemistry* (in the press).
13. Orly, J. & Schramm, M. *Proc. natn. Acad. Sci. U.S.A.* **72,** 3433–3437 (1975).
14. Shinitzky, M. & Inbar, M. *Biochim. biophys. Acta* **433,** 133–149 (1976).
15. Amdur, I. & Hammes, G. G. in *Chemical Kinetics: Principles and Selected Topics* (McGraw-Hill, New York, 1966).
16. North, A. M. in *The Collision Theory of Chemical Reactions in Liquids* (Methuen, London and Wiley, New York, 1964).
17. Hardt, S. & Cone, R. A. *Biophys. J.* (in the press).
18. Levitzki, A., Atlas, D. & Steer, M. L. *Proc. natn. Acad. Sci. U.S.A.* **71,** 2773–2776 (1974).
19. Hanski, E. & Levitzki, A. *Life Sci.* **22,** 53–60 (1978).
20. Poo, M. M. & Cone, R. A. *Nature* **247,** 438–440 (1974).
21. Axelrod, D. *et al. Proc. natn. Acad. Sci.* U.S.A. **73,** 4594–4598 (1976).
22. Dale, R. E., Chen, L. A. & Brand, L. *J. biol. Chem.* **252,** 7500–7510 (1977).
23. Cone, R. A. *Nature new Biol.* **236,** 39–43 (1972).
24. Klein, I., Moore, L. & Pastan, I. *Biochim. biophys. Acta* **506,** 42–53 (1978).
25. Houslay, M. D., Hesketh, G. A., Smith, G. B., Warren, G. B. & Metcalf, J. C. *Biochim. biophys. Acta* **436,** 495–504 (1976).
26. Warren, G. B., Toon, P. A., Birdsall, N. J. M., Lee, A. G. & Metcalf, J. C. *Biochemistry* **13,** 5501–5507 (1974).
27. Kimelberg, H. K. *Biochim. biophys. Acta* **413,** 143–156 (1975).
28. Thilo, L. Träuble, M. & Overath, P. *Biochemistry* **16,** 1283–1299 (1977).
29. Steer, M. L. & Levitzki, A. *J. biol. Chem.* **250,** 2080–2084 (1975).

Monoclonal antibody to a human histocompatibility alloantigen, HLA-A2

ANTIGENS of the HLA-A, -B, -C and -DR series have been defined mainly using sera from multiparous women[1], which generally are comparatively weak, serologically complex and available in only limited quantities. Furthermore, large screening programmes are required to find usable sera, especially for rare specificities. Although antisera to HLA determinants can be produced in high titres by immunising animals of various species with purified HLA-A, -B, -C, -D antigens[2-5], they usually require extensive absorption to reveal polymorphic specificity and so have not so far been used as routine tissue-typing reagents. The development of techniques for producing hybrid cell lines secreting monoclonal antibodies of defined specificity[6,7], provides a way of circumventing the problems of generating human polymorphic antisera by heteroimmunisation and has already been used to produce a monoclonal antibody against the human blood group A determinant[8]. Here we describe the production of a cell line secreting an anti-HLA-A2 antibody.

Female BALB/c mice were immunised subcutaneously with 20 μg of purified papain-solubilised HLA-A2 antigen[9] in complete Freund's adjuvant. After 1 month an equivalent boost in incomplete Freund's adjuvent was given and 2 weeks later the mice bled and their sera tested for anti-HLA-A, -B, -C antibodies. Serum titres of between 1 in 1,000 and 1 in 5,000 as measured by an indirect binding assay[10] on human B-cell lines expressing HLA-A, -B, -C antigens were observed. The titres against the B-cell line Daudi, which does not express HLA-A, -B, -C or the associated β_2 microglobulin (β_2m) antigens[11] were less than 50 and not significantly different from that obtained with pre-immune serum. No specificity for HLA-A2 could be detected by this method. The three mice with the highest serum titres were boosted intravenously with 10 μg of HLA-A2 antigen in saline, killed 5 days later, their spleens removed and the spleen cells fused with the mouse myeloma cell line P3-NSL/1-Ag4-1 (refs 6, 7, 12). After fusion the cells were divided between 192 two-ml culture wells. Visible colonies of hybrid cells were observed in all wells after 2 weeks and the culture supernatants were then tested for binding activity against an HLA-A2 positive B-cell line (Bri 8). Positive activity was seen in 109/192 supernatants, and the cultures from 39 that gave greater than 6 times the background were grown and their supernatants tested against Bri 8 and Daudi cell lines. This distinguished anti-HLA-A, -B, -C or β_2m antibodies from those directed against other cell surface components. Of the 39 cultures 38 showed no reaction with Daudi cells, thus indicating the high specificity of this procedure for generating antibodies to HLA-A, -B, -C or β_2m antigens. On continued growth, 14 cultures retained strong activity and were selected for further study. The binding activity of all 14 supernatants was fully inhibited by pre-incubation with purified HLA-A2 antigen, but only one of them could be inhibited by pure β_2m. Thus, the majority of the antibodies produced show specificity for the HLA-A, -B, -C chain, in contrast to one report suggesting a predominance of anti β_2m antibodies following immunisation with whole cells[13].

Of the 14 supernatants 10 reacted equally with all the cells from a panel of lines expressing a range of HLA-A, -B, -C antigenic specificities and so probably recognise antigenic determinants common to all HLA-A, -B, -C antigens comparable to a monoclonal antibody (W6/32) which has previously been described[18]. Polymorphic reaction patterns[4] were observed in 4 supernatants. PA2.1 reacted only with cells typed as HLA-A2 and did not react with cells expressing the highly cross-reactive HLA-A28 antigen[14] (Fig. 1).

Supernatants PA2.2, PA2.3 and PA2.4 reacted with some HLA-A2-negative cells as well as all HLA-A2-positive cells. These antibodies probably recognise cross-reacting antigenic determinants which HLA-A2 shares with some other HLA-A, -B, -C antigens. Their detailed specificity is being analysed, but

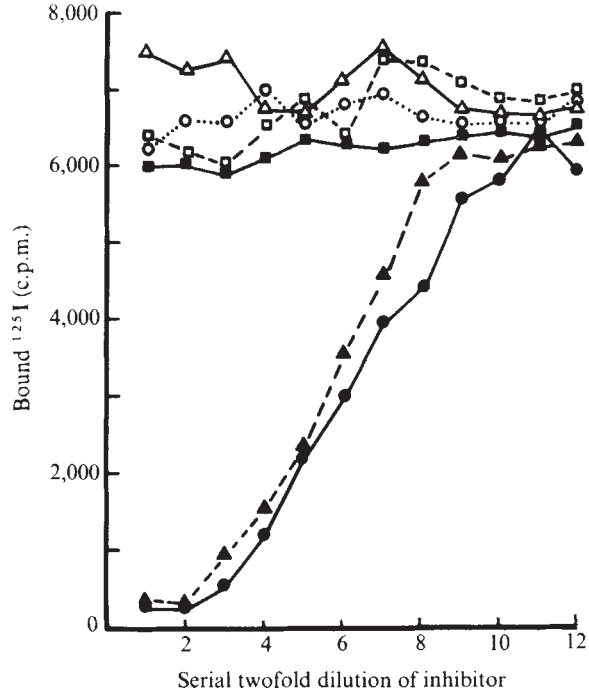
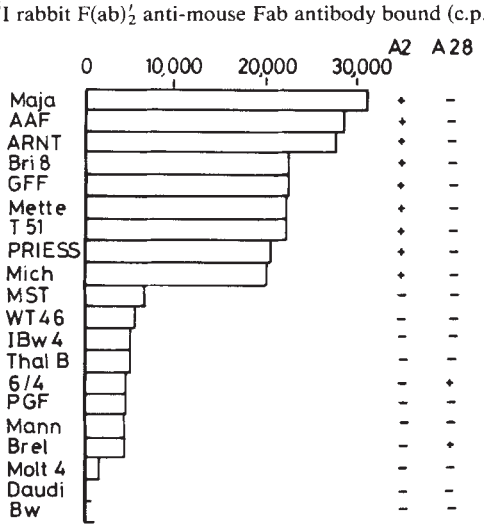


Fig. 2 Inhibition of PA2.1 antibody by purified papain-solubilised HLA-A, -B antigens with different antigenic specificity. HLA-A, -B antigen preparations (200 μg ml^{-1}) were serially diluted and 50 μl preincubated with 5 μl of PA2.1 supernatant for 1 h at 20 °C. Residual PA2.1 binding activity was then measured as for Fig. 1 with Bri 8 as the target cell (see refs 9 and 21). ●, HLA-A2 from J. Y. cells (homozygous *A2, B7*); ▲, HLA-A2 from 231 cells (homozygous *A2, B27*); □, HLA-A28 from L.B. cells (homozygous *A28, B40*); △, HLA-B40 from LB cells, ○, a mixture of HLA-A3, 25, B12, 27 from IM-1 cells (A3, 25, B12, 27); ■, a mixture of HLA-B7, 12 from RPMI 4265 cells (A2; B7, 12).

Fig. 1 Specificity of PA2.1 antibody for cell lines expressing HLA-A2 as measured by indirect trace binding assay. 10^6 cells were incubated with 50 μl of PA2.1 culture supernatant for 1 h at 4 °C, washed three times and then incubated with ^{125}I rabbit F(ab)$'_2$ anti-mouse Fab antibody, 300,000 c.p.m. for 1 h at 4 °C. The cells were washed four times and then counted for radioactivity. BW is a mouse thymoma cell line, Molt 4 a human T-cell line, all others are human B-cell lines[8].



© Macmillan Journals Ltd 1978

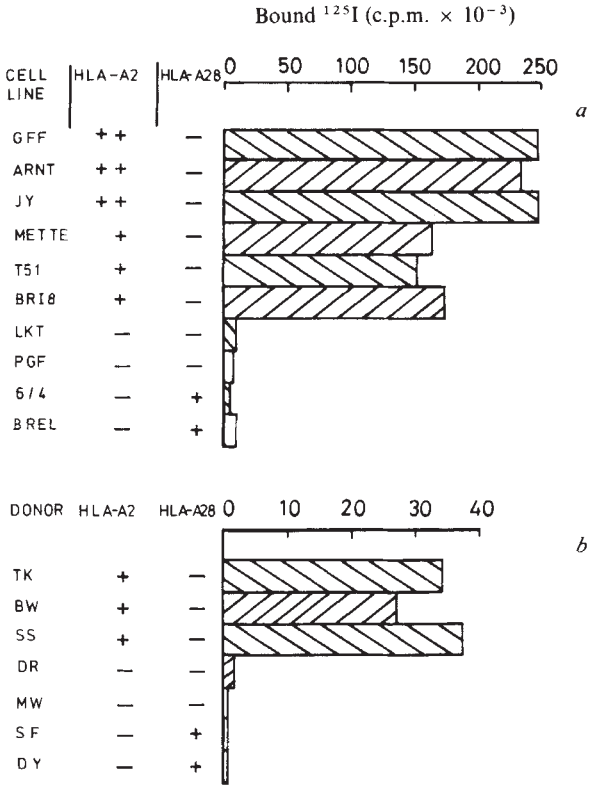


Fig. 3 Quantitation of HLA-A2 relative to total HLA-A, B, -C antigens on (*a*), B cell lines and (*b*), peripheral blood lymphocytes. The indirect binding of PA2.1 and W6/32 antibodies under saturating conditions of 1st antibody (20 μg ml^{-1}) and 2nd antibody (40 μg ml^{-1}) was measured on 5×10^5 cells in (*a*) and 2×10^6 cells in (*b*). The experimental procedure was as for Fig. 1 with an input of 3×10^6 counts per minute. To compare different cells within each group, the values plotted for the PA2.1 antibody binding have been normalised to a constant amount of radioactivity (representing total amount of HLA-A, -B, -C antigens) bound by the W6/32 antibody. This was the mean amount of radioactivity bound by the W6/32 antibody for all cells within a group and was 5×10^5 c.p.m. for B cell lines and 2.5×10^5 c.p.m. for peripheral blood lymphocytes. Cells are indicated as typed homozygous (++: based on family studies), heterozygous (+) and negative (−) for HLA-A2 and HLA-A28 by human alloantisera.

preliminary inhibition experiments with purified HLA-A, -B, -C antigens of different specificity indicate that supernatant PA2.2 probably reacts with a group including most HLA-A locus antigens.

Culture PA2.1 was cloned in soft agar[15] and one (PA2.1) of a number of colonies all with similar positive activity was arbitrarily selected for further study. The immunoglobulin secreted by PA2.1 was biosynthetically labelled with ^{14}C-lysine. Analysis by SDS-polyacrylamide gel electrophoresis, with other well characterised monoclonal antibodies as standards[16] showed that the PA2.1 antibody was an IgG molecule. It was not cytotoxic in complement-dependent antibody mediated cell lysis.

The specificity of the PA2.1 antibody for HLA-A2 was confirmed by inhibition with a variety of pure papain-solubilised HLA-A, -B, -C antigens (Fig. 2). Two preparations of HLA-A2 from different cell lines completely inhibited the PA2.1 antibody with similar specific activity, whereas four preparations representing eight other HLA-A, -B, -C antigenic specificities completely failed to inhibit. In particular, the PA2.1 antibody was not inhibited by the HLA-A28 preparation. It is unlikely that this was due to denaturation and loss of antigenic acitivity in the non-HLA-A2 antigen preparations, as all six fully inhibited the W6/32 monoclonal antibody, which recognises a determinant common to all HLA-A, -B, -C antigens that is dependent on the association of the HLA-A, -B, -C chain with β_2m and is easily inactivated by denaturation (data not shown).

HLA typing of B-cell lines often reveals patterns of serum cross reactions not usually observed with peripheral blood lymphocytes[17]. However, when PA2.1 antibody was tested by indirect binding against peripheral blood lymphocytes from a panel of 20 HLA-typed individuals it maintained its specificity for HLA-A2. Eight HLA-A2-positive individuals reacted strongly while 12 HLA-A2-negative individuals, including three who were HLA-A28 did not react (Fig. 3*b*). The same panel of peripheral lymphocytes was analysed in the fluorescence activated cell sorter after incubation with PA2.1 antibody and fluorescent labelled rabbit F(ab)$_2$ anti mouse Fab antibody, using W3/25 monoclonal antibody with specificity for rat lymphocytes as control[18]. There was again complete concordance with HLA-A2, confirming the specificity of PA2.1 antibody for HLA-A2 and showing its potential use for cell separation. When the indirect binding assay is used with saturating amounts of antibody it is possible to quantitate the numbers of antigenic sites per cell[10]. By comparing the binding of PA2.1 and W6/32 antibodies, the proportion of the total HLA-A, -B, -C antigen which is HLA-A2 can also be estimated. As shown in Fig. 3*a*, cell lines homozygous and heterozygous for HLA-A2 could readily be distinguished. On this basis HLA-A2 accounted for 40–50% of the total HLA-A, -B, -C in homozygous cells and 30–35% of the total in heterozygous cells. Peripheral lymphocytes (Fig. 3*b*) had, per cell, about one eighth of the amount of HLA-A, -B, -C antigens found on B-cell lines as measured by the W6/32 antibody, although within each cell type the expression of HLA-A, -B, -C antigens varied by a factor of about two. In contrast to the B-cell lines, HLA-A2 accounts for only 12–15% of the total HLA-A, -B, -C expressed as W6/32 reactivity on the peripheral lymphocytes of individuals heterozygous for HLA-A2. This difference in apparent relative expression of HLA-A2 between EBV-transformed B-cell lines and a predominantly T-cell population of normal lymphocytes is being investigated further.

Xenogeneic immunisation between distantly related species, such as mouse and man, does not, in general, readily lead to the production of allospecific antibodies, although recently Lampson *et al.*[19] have produced limited quantities of anti-HLA-A2 antibodies using a mouse spleen cell culture system. We have shown, however, that it is possible by immunisation of mice with pure HLA antigen preparations to produce stable hybrid cell lines secreting mouse monoclonal antibodies directed at human polymorphic histocompatibility antigenic determinants. In addition to providing tissue typing reagents, these antibodies will be invaluable for the purification, separation and immunochemical analysis of HLA-A, -B, -C, -D molecules. We are, for example, now using the PA2.1 antibody to determine the antigenic structure of HLA-A2 and compare it with HLA-A28. Serological analysis of cross-reacting monoclonal antibodies such as PA2.2 may clarify the antigenic relationships between various HLA alleles and provide information on the evolution of this complex genetic system[20].

We thank Harry Orr and Jack Strominger for HLA antigen preparations, Alan Williams for the W3/25 and rabbit F(ab)$_2$ antimouse Fab antibodies and Don Mason for cell sorter analyses. P. P. was supported by a Junior Fellowship from the Society of Fellows, Harvard University. We thank Colin Barnstable for many helpful discussions and Barbara Gunn and Beryl Winearls for their advice and technical help, and the MRC for support.

P. PARHAM
W. F. BODMER

Genetics Laboratory, University of Oxford,
South Parks Road,
Oxford, UK

Received 29 August; accepted 13 October 1978.

1. Bodmer, W. F., Batchelor, J. R., Bodmer, J. G., Festenstein, H. & Morris, P. J. *Histocompatibility Testing 1977* (Munksgaard, Copenhagen, 1978).
2. Metzgar, R. S. & Miller, J. L. *Transplantation* **13**, 467–471 (1972).
3. Sanderson, A. R. & Welsh, K. I. *Transplantation* **16**, 304–312 (1973).
4. Robb, R. J. *et al.* *Transplantation* **19**, 445–447 (1975).

5. Barnstable, C. J. thesis, Univ. Oxford (1978).
6. Köhler, G. & Milstein, C. *Nature* **256**, 495–497 (1975).
7. Köhler, G. & Milstein, C. *Eur. J. Immun.* **6**, 511–519 (1976).
8. Barnstable, C. J. *et al. Cell* **14**, 809-820 (1978).
9. Parham, P. *et al. J. biol. chem.* **252**, 7555–7562 (1977).
10. Williams, A. F. *Contemp. Top. molec. Immun.* **6**, 83–116 (1977).
11. Arce-Gomez, B., Jones, E. A., Barnstable, C. J., Solomon, E. & Bodmer, W. F. *Tissue Antigens* **11**, 96–112 (1978).
12. Köhler, G., Howe, S. C. & Milstein, C. *Eur. J. Immun.* **6**, 292–295 (1976).
13. Trucco, M. M., Stocker, J. W. & Ceppellini, R. *Nature* **273**, 666–668 (1978).
14. Kissmeyer-Nielsen, F., Svejgaard, A. & Hauge, M. *Nature* **219**, 1116–1119 (1968).
15. Cotton, R. G. H., Secher, D. S. & Milstein, C. *Eur. J. Immun.* **3**, 136–140 (1973).
16. Köhler, G., Hengartner, H. & Shulman, M. J. *Eur. J. Immun.* **8**, 82–88 (1978).
17. Dick, H. M. *et al. Histocompatibility Testing 1975* 671–676 (Munksgaard, Copenhagen, 1975).
18. Williams, A. F., Galfre, G. & Milstein, C. *Cell* **12**, 663–673 (1977).
19. Lampson, L. A. *et al. Nature* **271**, 461–462 (1978).
20. Bodmer, W. F. *Nature* **237**, 139–145 (1972).
21. Parham, P. *et al. Proc. natn. Acad. Sci. U.S.A.* **72**, 1594–1598 (1975).

Effect of polarisation of horizontal cells on the on-centre bipolar cell of carp retina

THERE are two types of bipolar cell in the vertebrate retina, the on-centre and the off-centre cells[1,2]. The on-centre bipolar cells respond with depolarisation to the light spot in the centre of the receptive field and with hyperpolarisation to surround illumination. It is generally assumed that the effect of surround illumination is mediated by horizontal cells, as negative feedback from horizontal cells to photoreceptors has been demonstrated in the turtle retina[3]. Here we have attempted to clarify the role of horizontal cells in surround antagonism by hyperpolarising horizontal cells by applying extrinsic current in the retina of the carp, *Cyprinus carpio*. The response of the bipolar cells was similar to that elicited by surround illumination. However, the ionic mechanisms differed depending on the morphological type[4] of horizontal cell. An analysis of these mechanisms has led us to conclude that the external horizontal cells receiving inputs exclusively from cones modify the cone–bipolar transmission and the intermediate horizontal cells receiving inputs exclusively from rods modify the rod–bipolar transmission.

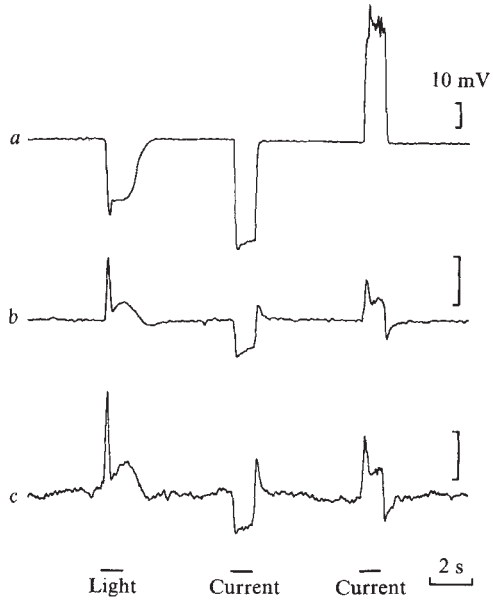


Fig. 1 Effect of polarisation of an intermediate horizontal cell on a bipolar cell. *a*, Potential changes of the intermediate horizontal cell; *b*, responses of the bipolar cell recorded simultaneously with *a*; *c*, responses of the same bipolar cell during hyperpolarisation of its membrane by steady current of 2.3 nA. In each trace the response to diffuse white light was tested first (left), then the horizontal cell was polarised in the dark to both hyperpolarising (middle) and depolarising (right) directions by extrinsic current through one barrel of a double-barrelled electrode. Timing of light and current stimuli is indicated below. The current used to polarise the horizontal cell was 10 nA. The coupling resistance of the double-barrelled electrode was 0.9 MΩ when measured in Ringer solution. The voltage drop across the coupling resistance was not compensated in these experiments. Calibration of 10 mV is shown at the end of each trace.

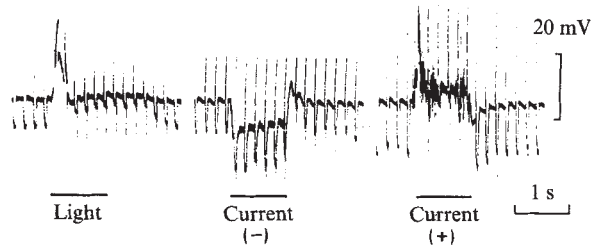


Fig. 2 Bridge records of bipolar cell responses to light (left) and to hyperpolarisation (middle) and depolarisation (right) of an intermediate horizontal cell by extrinsic current. Negative current pulses of 2.4 nA and of 50 ms duration were applied into the bipolar cell at a frequency of 5 pulses s^{-1} through the bridge circuit. The bridge was adjusted so that the decrease in the membrane resistance was indicated by a decrease in the height of pulses deflecting downward. The 20-mV change in the pulse height corresponds to 8.3-MΩ change in the membrane resistance. The response to light is accompanied by a resistance decrease. The hyperpolarising response to polarisation of the horizontal cell is accompanied by an increase and the depolarising response by a decrease in the membrane resistance.

Cajal[4] has described three morphologically different types of horizontal cell in the teleost fish retina; external, intermediate and internal. Both morphological and physiological studies have shown that external horizontal cells receive inputs exclusively from cones and intermediate horizontal cells exclusively from rods[2,5,6]. Recent studies further revealed two sublayers of external horizontal cells[7,8]. The cells in the distal layer are not colour-coded and give the response known as the L-type. Those in the second layer are colour-coded and give the response known as the C-type. The internal horizontal cells of Cajal are now considered as the processes of external horizontal cells[9]. These different types of horizontal cells were polarised in the present experiments by extrinsic current and its effect on bipolar cells was studied. The bipolar cells studied in these experiments were on-centre cells, because they were most frequently recorded and their responses were relatively stable. The bipolar cells recorded were identified by staining with Procion yellow as the large bipolar cells of Cajal which are known to have connections with both rods and cones[5,10].

Single- and double-barrelled microelectrodes, which could be advanced independently, were set almost parallel with their tip distance within 50 μm on the horizontal plane and were inserted into the isolated inverted retina vertically from above. When the responses of a bipolar and a horizontal cell were recorded, by single- and double-barrelled electrodes respectively, their cell types were identified by the response to diffuse white light of about 0.4 lm cm^{-2} and by changing patterns and wavelengths of the light stimulus if necessary. After confirmation of the response types, an extrinsic current of the order of 10 nA was passed through one barrel of the double-barrelled electrode in the absence of light but at the same timing as the light stimulus. In order to examine the ionic mechanisms of bipolar cell responses elicited by polarisation of horizontal cells, the membrane potential of bipolar cells was also changed by a steady polarising current or a train of current pulses of the order of 2 nA through a bridge circuit built in a preamplifier (WPI M701R). The steady current was used to estimate the reversal potential and the pulse current to measure the membrane resistance changes.

An example of simultaneous recordings from an intermediate horizontal cell (*a*) and an on-centre bipolar cell (*b*) is shown in

0028–0836/78/0276—0399$01.00
© Macmillan Journals Ltd 1978

Fig. 1. The response of intermediate horizontal cells can be easily distinguished from that of external horizontal cells by its time course. The hyperpolarisation of intermediate horizontal cells by current elicited a hyperpolarising response in the on-centre cell and the depolarisation a depolarising response. When the bipolar cell was hyperpolarised by a steady extrinsic current, both hyperpolarising and depolarising responses elicited by the horizontal cell polarisation were augmented (*c*). The depolarisation of the bipolar cell resulted in a decrease in the response amplitude. The measurement of the membrane resistance changes revealed that the hyperpolarising response was accompanied by an increase and the depolarising response by a decrease in the membrane resistance (Fig. 2). It is inferred from these results that the effect of polarisation of intermediate horizontal cells is mediated by ionic channels such as sodium channels which have a positive reversal potential. These channels are made more permeable by depolarisation and less permeable by hyperpolarisation of intermediate horizontal cells.

An example of simultaneous recordings from an L-type external horizontal cell (*a*) and an on-centre bipolar cell (*b*) is shown in Fig. 3. The hyperpolarisation of external horizontal cells also elicited a hyperpolarising response in the on-centre cells and the depolarisation a depolarising response. These responses were made smaller or inverted by hyperpolarisation of the bipolar cell membrane by steady extrinsic current (*c*). They were augmented by depolarisation of the bipolar cell. The measurement of the membrane resistance changes revealed that the hyperpolarising response was accompanied by a decrease and the depolarising response by an increase in the membrane resistance (Fig. 4). It seems therefore that the effect of polarisation of external horizontal cells is mediated by ionic channels which have a more negative reversal potential than the membrane potential in the dark like potassium and/or chloride channels, commonly observed in the inhibitory synapses. These channels are made more permeable by hyperpolarisation and less permeable by depolarisation of the external horizontal cells.

Studies on the synaptic transmission from photoreceptors to on-centre bipolar cells indicate that rod signals are mediated by a permeability increase of subsynaptic sodium channels in the light and cone signals by a permeability decrease of subsynaptic potassium and/or chloride channels[11]. Our results are consistent with these data if we assume that intermediate horizontal cells

Fig. 3 Effect of polarisation of an external horizontal cell on a bipolar cell. *a*, Potential changes of the external horizontal cell; *b*, responses of the bipolar cell recorded simultaneously with *a*; *c*, responses of the same bipolar cell during hyperpolarisation of its membrane by current. Recording conditions were the same as in Fig. 1 except that the current used to polarise the horizontal cell was 13 nA.

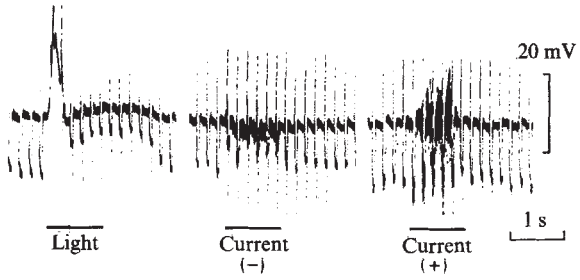


Fig. 4 Bridge records of bipolar cell responses to light (left) and to polarisation of an external horizontal cell by current (middle and right). Conditions of the bridge records were the same as in Fig. 2. The relatively large amount of current (20 nA) necessary to polarise the external horizontal cell generated an electrode noise, which was also picked up capacitatively in the bipolar cell records. Although the bridge records shown here are partly masked by such noise, it is seen that the hyperpolarising response to polarisation of the horizontal cell is accompanied by a decrease and the depolarising response by an increase in the membrane resistance.

not only receive exclusive inputs from rods and selectively affect the rod–bipolar transmission in a way expected from the feedback hypothesis, but the external horizontal cells receive exclusive inputs from cones and selectively affect the cone–bipolar transmission.

The effect of polarisation of intermediate horizontal cells on the bipolar cells were clearly observed in all of the 11 pairs studied. On the other hand, the effect of polarisation of L-type external horizontal cells were observed in 5 out of more than 20 pairs studied. The effect was not prominent even in these 5 units. One reason for such results is that the input resistance of external horizontal cells is generally low which makes it difficult to polarise them sufficiently to affect the bipolar cells because of the restriction on the current flowing through the fine electrodes. It is also probable that the large bipolar cells receive a relatively small number of inputs from cones which mediate this effect.

The effect of polarisation of C-type horizontal cells on the on-centre bipolar cells was also studied in several pairs. But, so far we have been unable to detect any changes in the bipolar cells. It has been reported from morphological studies on goldfish retina that the C-type cells have direct connections with either blue or green cones but not with rods or red cones[12], and that a certain type of large bipolar cell has connections with rods and red cones but not with blue or green cones[13]. Although it is difficult to exclude the possibility that the current passed is not large enough to polarise the cell because of its low input resistance, the results probably indicate that the C-type cells do not affect the type of on-centre bipolar cells studied because they do not share the common photoreceptors.

This work was partly supported by grants from the Ministry of Education of Japan and by grant EY-02392 from the National Eye Institute, USPHS.

JUN-ICHI TOYODA
KEIICHI TONOSAKI

Department of Physiology,
St. Marianna University School of Medicine,
Sugao, Takatsu-ku, Kawasaki,
213 Japan

Received 7 July; accepted 21 September 1978.

1. Werblin, F. S. & Dowling, J. E. *J. Neurophysiol.* **32,** 339–355 (1969).
2. Kaneko, A. *J. Physiol., Lond.* **207,** 623–633 (1970).
3. Baylor, D. A., Fuortes, M. G. F. & O'Bryan, P. M. *J. Physiol., Lond.* **214,** 265–294 (1971).
4. Cajal, S. R. *La Cellule* **9,** 119–253 (1892).
5. Stell, W. K. *Am. J. Anat.* **121,** 401–424 (1967).
6. Kaneko, A. & Yamada, M. *J. Physiol., Lond.* **227,** 261–273 (1972).
7. Mitarai, G., Asano, T. & Miyake, Y. *Jap. J. Opthal.* **18,** 161–176 (1974).
8. Hashimoto, Y., Kato, A., Inokuchi, M. & Watanabe, K. *Vision Res.* **16,** 25–29 (1976).
9. Stell, W. K. *J. comp. Neurol.* **159,** 503–520 (1975).
10. Scholes, J. H. *Phil. Trans. R. Soc.* **270,** 61–118 (1975).
11. Saito, T., Kondo, H. & Toyoda, J. *Vision Res.* **18,** 591–595 (1978).
12. Stell, W. K. & Lightfoot, D. O. *J. comp. Neurol.* **159,** 473–502 (1975).
13. Stell, W. K., Ishida, A. T. & Lightfoot, D. O. *Science* **198,** 1269–1271 (1971).

Change in synaptic channel gating during neuromuscular development

FORMATION of neuromuscular synapses involves accumulation of acetylcholine receptors (AChRs) in the postsynaptic membrane[1-5]. The ionic channels associated with synaptic AChRs differ from the AChR channels present in the surface membrane of non-innervated muscle fibres. In particular, the mean open time of synaptic channels is much shorter than that of non-synaptic channels[6-10]. The question therefore arises at which stage of neuromuscular development the gating properties of synaptic channels become similar to those observed in adult junctions. The experiments described here show a difference in the mean open time of channels in early neonatal and adult rat endplates and suggest that a change in gating behaviour takes place during postnatal development.

Experiments were performed at 20–23°C on isolated soleus and omohyoideus muscles taken from newborn rats between 1 and 18 d of age. Endplates were voltage clamped to −70 mV with a conventional two-microelectrode clamp. Acetylcholine (ACh) was released iontophoretically to induce steady endplate currents (e.p.cs) of 20–50 nA average amplitude. Channel mean open times were estimated from the autocorrelation function (a.c.f.) of fluctuations of the e.p.c. around its mean value. Assuming a simple two-state model of channel gating[6,11,12], the mean channel open time is given by the value of the decay time constant τ of the a.c.f.[7,12].

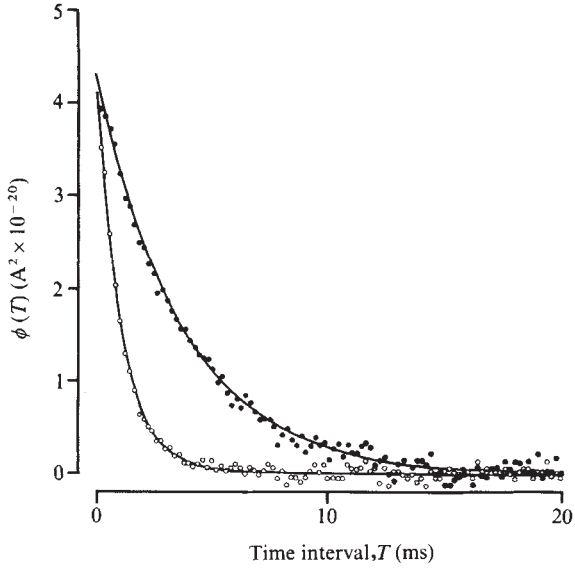


Fig. 1 Autocorrelation functions of ACh-induced current fluctuations recorded from voltage-clamped endplates of rat soleus muscle on postnatal days 8 (●) and 18 (○). Data points represent value of autocorrelation function $\phi(T)$ for different time intervals T. Mean ACh-induced currents were $I = 25$ nA and $I = 22$ nA, respectively. Membrane potential was clamped to −70 mV; temperature 22°C. Clamp currents before and during application of ACh were recorded on magnetic tape at low and high gain. The high gain record was bandpass filtered (1 Hz–2.5 kHz, 48 dB per octave roll off). High gain records were digitised at 200μs intervals and their a.c.f. calculated on a SAICOR auto-cross correlation analyser. Typically current samples of 6–12 s duration were analysed. a.c.fs of control and of ACh-evoked current samples were stored in a PDP 11/34 computer. The a.c.f. of the control sample was subtracted from the a.c.f. of the ACh-induced current sample. The resulting data points were displayed on an oscilloscope screen together with a machine-generated exponential (solid lines) corresponding to the expression $\phi(T) = \sigma^2 \cdot \exp(-T/\tau)$ where σ^2 represents the total variance of current fluctuations, and τ the a.c.f. decay time constant. For the a.c.fs shown, the curve-fitting yielded values of 3.9 ms (●) and 1.0 ms(○) for the decay time constant τ. Estimates of single channel conductances γ were obtained according to $\gamma = (\sigma^2/I)/(V - V_{rev})$ where V is the holding potential and V_{rev} the reversal potential (assumed to be 0 mV). The value of σ^2 is obtained from the extrapolated ordinal intercept of the fitted exponential. The single channel conductances were estimated as $\gamma = 25$ pS (day 8) and $\gamma = 27$ pS (day 18). All experiments were carried out in a solution containing (mM): NaCl, 140; KCl, 2; $MgCl_2$, 5; KH_2PO_4, 1; $NaHCO_3$, 12; pH 7, 2, oxygenated with 95% O_2 and 5% CO_2.

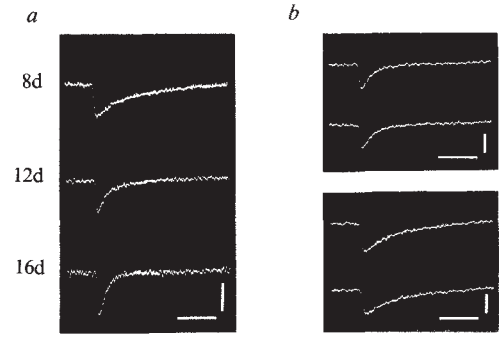


Fig. 2 Time course of m.e.p.cs at different stages of postnatal development. *a*, Change of soleus m.e.p.c. decay time course between postnatal days 8 and 16. Representative examples of single m.e.p.cs at different days after birth are shown. Note complex decay of m.e.p.cs on postnatal day 12. Recordings were made from soleus muscles of littermates. Membrane potential was clamped to −70 mV, 21–23°C. m.e.p.cs were recorded on magnetic tape and were replayed into a PDP-11/34 computer at a sampling rate of 20 kHz. Individual m.e.p.cs were displayed on an oscilloscope screen together with either one or the sum of two machine-generated exponentials which were fitted to the m.e.p.c. decay by eye. In the case of m.e.p.cs with doubly exponential decays one exponential was first fitted to the m.e.p.c. tail and subsequently another one to the difference between the data points and the extrapolated slower exponential. The mean $\bar{\tau}$ values in each fibre were on day 8: 4.6 ± 0.3 ms; on day 12: 1.1 ± 0.1 ms for the fast and 4.5 ± 0.3 ms for the slow component of m.e.p.c. decay; and on day 16: 1.2 ± 0.2 ms. Mean ±s.e.m., values from 10 individually measured m.e.p.cs were averaged in each fibre. Similar results were obtained with two other litters. Staining of endplate acetylcholinesterase activity by the Karnovsky method[17] did not show in the light microscope any obvious change in the staining intensity during this period. Calibration bars represent 4 ms and 2.5 nA. *b*, Recordings of m.e.p.cs on postnatal day 8 from neuromuscular junctions of omohyoideus (upper two traces) and soleus (lower two traces) muscle of the same rat. The average values of m.e.p.c. decay time constants were $\bar{\tau} = 1.3 \pm 0.1$ ms for the omohyoideus endplate and $\bar{\tau} = 4.3 \pm 0.2$ ms for the soleus endplate, averaging 10 individual m.e.p.cs in each experiment. Calibration bars represent 2 ms and 2 nA.

Figure 1 shows the a.c.fs of ACh-evoked e.p.c. fluctuations recorded from voltage clamped soleus endplates on postnatal days 8 and 18. The a.c.f. decay was singly exponential in both experiments but the decay time constants differed considerably. The value of τ on day 8 was 3.9 ms, which is about four times longer than the value of 1.0 ms found at the day 18 endplate. In seven experiments carried out before day 8, τ was 3.6 ± 0.2 ms (mean ±s.e.m.). This is comparable to the value found for channels in the extrasynaptic membrane of adult denervated[8] or early neonatal fibres (B.S. and H.R.B., unpublished) and for synapses formed *in vitro*[13]. Conversely, the mean channel open time observed in endplates on day 18 had an average value of $\tau = 1.0 \pm 0.1$ ms (four endplates), which is identical to that found in adult endplates[8,10,14]. From around postnatal day 8 to around day 17, however, the a.c.f. decay could be fitted adequately only by the sum of a fast and a slow exponential. Their average time constants were $\tau_{fast} = 1.0 \pm 0.1$ ms and $\tau_{slow} = 3.9 \pm 0.1$ ms (seven endplates). These are very similar to the time constants found at adult (≥18 d) and early neonatal (≤8 d) endplates, respectively. The amplitude of the fast component increased

0028–0836/78/0276—0401$01.00
© Macmillan Journals Ltd 1978

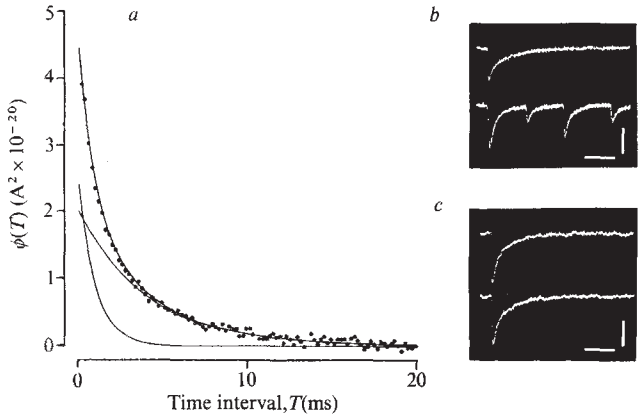


Fig. 3 Presence of two channel populations with different mean open times in neonatal omohyoideus endplate membrane. *a*, Autocorrelation function of endplate current fluctuations induced by iontophoretic ACh application on postnatal day 2. Data points represent value of the a.c.f. $\phi(T)$ for different time intervals T. The sum of two single exponentials, each of the form given in Fig. 1, was fitted to the data points (line drawn through data points). The two other lines represent the single exponentials. The time constants were $\tau_{fast} = 1.1$ ms for the fast and $\tau_{slow} = 3.9$ ms for the slow exponential. The ratio of their variances is $\sigma^2_{fast}/\sigma^2_{slow} = 1.18$. The slow exponential was first fitted to the tail of the a.c.f. decay, and subsequently the fast exponential was fitted to the difference between the data points and the extrapolated slow exponential. Membrane potential: −70 mV; 23°C. Mean ACh-induced endplate current was 21 nA. *b*, Recordings of m.e.p.cs on the same fibre as in *a*. m.e.p.c. decay is split into fast and slow component. The decay time course can be fitted by the sum of two exponentials with the time constants $\bar{\tau}_{fast} = 1.2$ ms and $\bar{\tau}_{slow} = 4.4$ ms. Mean values from 20 individually measured m.e.p.cs. *c*, Focal extracellular recordings of m.e.p.cs on postnatal day 4 from a voltage-clamped fibre (−70 mV, 22 °C). m.e.p.c. decay shows clear splitting in two decay phases and can be fitted by the sum of two exponentials. The respective time constants are $\bar{\tau}_{fast} = 1.0$ ms and $\bar{\tau}_{slow} = 5.3$ ms. The calibration bars represent 5 ms and 1 nA for *b* and 5 ms and 0.5 mV for *c*.

gradually during this transition period, whereas the amplitude of the slow component decreased correspondingly suggesting that from postnatal day 8 onwards the number of endplate channels with a short mean open time increases up to postnatal day 18 when all channels have fast kinetics.

Mean channel open time of junctional AChR channels can also be estimated from the time course of miniature endplate currents (m.e.p.cs). The falling phase of an m.e.p.c. in adult endplates is a single exponential the time constant of which, $\bar{\tau}$, approximately equals the mean open time of ACh-activated channels[11,15]. Developmental changes in the mean open time of junctional AChR channels should therefore be reflected in the m.e.p.c. decay time course. This is illustrated in Fig. 2*a* which shows examples of m.e.p.cs recorded from soleus endplates at various times after birth. To postnatal day 8 the m.e.p.c. decay is singly exponential with a mean value for $\bar{\tau}$ of 4.7 ± 0.2 ms (five endplates). This is four to five times greater than the value reported for adult endplates[14]. From around day 8 to around day 17, m.e.p.cs decayed with two time constants. From day 18 onwards m.e.p.c. decays were singly exponential again but with a much shorter time constant than is found in early neonatal fibres, the average value of $\bar{\tau}$ being 1.2 ± 0.1 ms (four endplates). During the transition period the m.e.p.c. decay could be satisfactorily fitted by the sum of two exponentials. Their time constants were similar to those of m.e.p.c. decays observed in early neonatal and adult endplates, respectively. The fast component becomes relatively more pronounced with age. For example, comparing measurements at four endplates on days 9 and 17 the average amplitude ratio of the fast to the slow component increased from 0.34 to 3.4.

A similar change in mean channel open time was observed in neonatal omohyoideus muscle, but it occurred earlier in postnatal development than in the soleus muscle. This is illustrated in Fig. 2*b*, which shows m.e.p.c. recordings from endplates in the omohyoideus and soleus muscles of the same rat at day 8. m.e.p.cs of the omohyoideus muscle decayed much faster than m.e.p.cs recorded from the soleus endplate. Their $\bar{\tau}$ value is comparable to that of m.e.p.cs in adult fibres already at this age. To determine the onset of the transition in channel gating in omohyoideus muscle, m.e.p.cs were recorded on postnatal days 1 and 2. In three out of four endplates studied on day 1 and in all four fibres studied on day 2, m.e.p.cs with doubly exponential decays were observed (Fig. 3*b*, *c*). The time constant of the fast component was $\bar{\tau}_{fast} = 1.2 \pm 0.1$ ms (seven endplates) comparable to the $\bar{\tau}$ value of m.e.p.cs in adult fibres. The average ratio of the slow to fast time constant was 4.3. Similarly the a.c.f. of ACh-induced e.p.c. fluctuations is fitted best by a double exponential decay during the first postnatal week (Fig. 3*a*) with the average time constants $\tau_{fast} = 1.2 \pm 0.1$ ms and $\tau_{slow} = 4.6 \pm 0.2$ ms (four endplates). From day 8 onwards only a fast component was resolved in both a.c.f. and m.e.p.c. decays, their respective time constants being $\tau = 1.0 \pm 0.08$ ms (four endplates) and $\bar{\tau} = 1.3 \pm 0.1$ ms (six endplates). This shows that synaptic channels in omohyoideus muscle have fast kinetics characteristic of the adult endplate, earlier during development than do those of soleus endplates. The fact that m.e.p.c. decays showed a fast component on postnatal day 1 suggests that in omohyoideus muscle a considerable fraction of adult-type channels is already present at birth.

Assuming neuromuscular synaptogenesis in rat begins at day 15 of gestation in the omohyoideus and at day 17 in soleus muscle[16], the experiments would indicate that a conversion of channels from an 'embryonic' to the adult type occurs in omohyoideus muscle between 6 and 14 d and in soleus between 13 and 22 d after the initial signs of neuromuscular transmission. It might be asked what influence causes the change of synaptic channel gating properties at a particular stage of development and why there is a difference between the omohyoideus and the soleus muscle? Neck muscles become active before the muscles of the lower extremities[16] and it is possible that earlier usage of neck muscles, for example during suckling, might cause this difference. Alternatively, maturation of neuromuscular junctions might be slower in red than in white muscle fibres.

We thank Dr J. Patlak for comments on the manuscript. H.R.B. was supported by the Swiss NSF.

B. SAKMANN

Abteilung Neurobiologie,
Max-Planck-Institut für biophysikalische Chemie,
3400 Göttingen-Nikolausberg, FRG

H. R. BRENNER

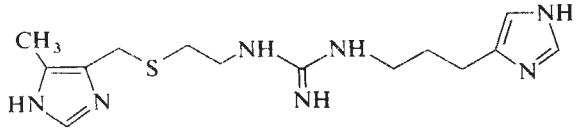
Physiologisches Institut der Universität Basel,
CH-4051 Basel, Vesalgasse 1, Switzerland

Received 24 July; accepted 2 October 1978.

1. Lomo, T. & Slater, C. R. *J. Physiol., Lond.* **275**, 391–402 (1978).
2. Bevan, S. & Steinbach, J. H. *J. Physiol., Lond.* **267**, 195–213 (1977).
3. Anderson, M. J. & Cohen, M. W. *J. Physiol., Lond.* **268**, 751–773 (1977).
4. Brenner, H. R. & Sakmann, B. *Nature* **271**, 366–368 (1978).
5. Fischbach, G. D., Berg, D. K., Cohen, S. A. & Frank, E. *Cold Spring Harb. Symp. quant. Biol.* **40**, 347–357 (1976).
6. Katz, B. & Miledi, R. *J. Physiol., Lond.* **224**, 665–699 (1972).
7. Neher, E. & Sakmann, B. *J. Physiol., Lond.* **258**, 705–730 (1976).
8. Sakmann, B. *Fedn. Proc.* **37**, 2654–2659 (1978).
9. Dreyer, F., Walther, Chr. & Peper, K. *Pflügers Arch. ges. Physiol.* **366**, 1–9 (1976).
10. Dreyer, F., Müller, K. D., Peper, K. & Sterz, R. *Pflügers Arch. ges. Physiol.* **367**, 115–122 (1976).
11. Anderson, C. R. & Stevens, C. F. *J. Physiol., Lond.* **235**, 655–691 (1973).
12. Stevens, C. F. *Biophys. J.* **12**, 1028–1047 (1972).
13. Schuetze, S. M., Frank, E. F. & Fischbach, G. D. *Proc. natn. Acad. Sci. U.S.A.* **75**, 520–523 (1978).
14. Colquhoun, D., Large, W. A. & Rang, H. P. *J. Physiol., Lond.* **266**, 361–395 (1977).
15. Magleby, K. L. & Stevens C. F. *J. Physiol., Lond.* **223**, 173–174 (1972).
16. Angulo y Gonzales, A. W. *J. comp. Neurol.* **55**, 395–442 (1932).
17. Karnovsky, M. R. *J. Cell Biol.* **23**, 217–222 (1964).

Impromidine (SK&F 92676) is a very potent and specific agonist for histamine H_2 receptors

PREVIOUSLY described selective agonists for histamine H_2 receptors are 4-methylhistamine[1], related congeners[2-6] and dimaprit[7]; these compounds are generally less potent than histamine. We have now found that *N*-[3-(imidazol-4-yl)-propyl]-*N'*-{2-[(5-methylimidazol-4-yl)methylthio]ethyl}-guanidine (impromidine, SK&F 92676 (I)) is extremely potent as an H_2-receptor agonist and briefly report its properties here.



I

The action of impromidine was studied *in vitro* on two tissues known to contain histamine H_2 receptors[1], the guinea pig right atrium and the rat uterus; the results are summarised in Table 1. On the atrium, impromidine had 48 times the potency of histamine whereas on the uterus it was 9.3 times as potent. The maximum responses achievable by the two agonists were compared by construction of complete cumulative dose–response curves (Fig. 1). On the guinea pig atrium impromidine achieved the same maximum response as histamine but on the isolated rat uterus the maximum response relative to histamine was only 80%, clearly suggesting that the compound has a lower efficacy than histamine. Cimetidine[8] (a histamine H_2-receptor antagonist) competitively inhibited the action on the guinea pig atrium, giving a dose-related parallel displacement of the dose–response curve to the right without depressing the maximum response (see Table 1 legend). In the absence of propranolol, 100 μM cimetidine completely suppressed the response to impromidine (in concentrations of up to 1 μM) on the atrium, suggesting that it had no significant agonist interaction with adrenergic β receptors.

On the atropinised guinea pig isolated ileum, a tissue known to contain H_1 receptors, impromidine failed to exert an agonist effect at concentrations up to 10 mM. This indicates that it has <0.001% the agonist activity of histamine at H_1 receptors. In fact, impromidine competitively antagonised the agonist effect of histamine on the ileum ($pA_2 = 5.47$; Table 1). Impromidine also failed to stimulate contractions of the ileum in the absence of atropine, indicating that it is not a muscarinic agonist. Indeed, impromidine competitively antagonised the agonist action of carbachol on this preparation ($pA_2 = 4.89$; Table 1).

Fig. 1 Isolated guinea pig right atrium; cumulative dose–response curves to histamine (●) and SK&F 92676 (○). Each point is the mean of 10 observations ± s.e.m. 0.5 μM propranolol was present in the bathing fluid.

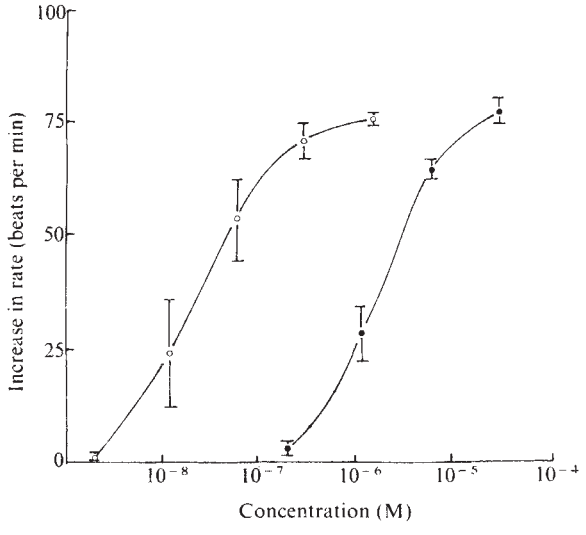


Table 1 Impromidine agonist activities, relative to histamine = 1

Preparation	Agonist activity mean (with 95% confidence limits)	% Maximum of histamine response (mean ± s.e.m.)	*n*
Guinea-pig ileum*	$<10^{-5}$		
Guinea-pig atrium†‡	48.1(37.6–60.1)	98.7 ± 3.6	10
Rat uterus†	9.3(5.7–15.2)	80.2 ± 2.6	22
Rat gastric acid secretion	16.8(11.3–27.1)	ND	16
Cat gastric acid secretion	~27	114 ± 18	5
Heidenhain pouch dog	~22	118 ± 11	5

SK&F 92676 was used as the trihydrochloride, a colourless crystalline solid, $C_{14}H_{23}N_7S$, 3HCl, of molecular weight 430.06 and melting point 198–200°C. It is very soluble in water and posesses three basic groups; a 3×10^{-3} molar solution (0.13% w/v) gives a *p*H of approximately 4.5. At this *p*H, or at *p*H 7.4, solutions are stable for at least 30 d at 20°C. For the *in vivo* studies, solutions were made up in normal saline. Rat gastric acid secretion data were obtained using lumen-perfused stomach of the anaesthetised, vagotomised rat[7,9] in a 2+2 Latin square assay. Cat gastric acid secretion was obtained from anaesthetised cats fitted with an acute gastric fistula and gastric juice was collected by drainage under gravity[7]. Dose–response curves to intravenous infusions of the agonist were constructed as described for the dog in Fig. 2. Conscious dogs were fitted with Heidenhain pouches which had been prepared 1–3 yr previously. Dose–response curves were constructed as described in Fig. 2. ND, not determined.

* Agonist assay on the ileum in the presence of 1 μM atropine[7]. Other actions: H_1-receptor antagonism of histamine, regression of log (DR–1) on log (concentration of SK&F 92676) ($n = 12$) had a slope of 0.95 ± 0.18 and gave $pA_2 = 5.47$ (95% confidence limits 5.16–5.74). Antagonism of carbachol, regression of log (DR–1) on log (concentration of SK&F 92676) ($n = 12$) had a slope of 0.95 ± 0.26 and gave $pA_2 = 4.89$ (95% limits 4.18–5.41).

† Comparative assays in the presence of 0.5 μM propranolol using a 2+2 Latin square design[7]. The ED_{50} values for SK&F 92676 were, respectively, 0.025 μM and 0.18 μM on atrium and uterus.

‡ Cimetidine $pA_2 = 6.53$ (6.09–6.98) ($n = 15$), not significantly different from the value against histamine[8] ($pA_2 = 6.10$(6.04–6.17)). The slope of the regression of log (DR–1) on log (concentration of cimetidine) = 0.81 ± 0.21, not significantly different from unity.

In vivo, impromidine was found to be a highly potent stimulant of gastric acid secretion (Table 1). Compared with histamine, using the technique of the lumen-perfused stomach of the anaesthetised, vagotomised rat[7,9], impromidine had 16.8 times the potency of histamine as a gastric acid secretagogue. The time course of the agonist response to impromidine after rapid intravenous injection was similar to that obtained to histamine. The infusion rate for near maximal acid secretion was 0.015 μmol per kg per min (compared with histamine, 0.25 μmol per kg per min). In the anaesthetised cat, impromidine was ~27 times as potent as histamine in stimulating gastric acid secretion, based on a comparison of the infusion rates producing 50% of the maximal response. As found with dimaprit[7], the maximal response seemed to be greater than that

Table 2 Blood pressure responses to doses of impromidine maximal for gastric acid secretion

Species	Dose of SK&F 92676 (μmol kg^{-1} min^{-1})	Mean fall in blood pressure (mm Hg)	Mean increase in pulse pressure (mm Hg)
		($n = 4$)	
Rat	0.015	17	ND
Cat	0.0032	27	20
Dog	0.0008	17	14

0028-0836/78/0276—0403$01.00

© Macmillan Journals Ltd 1978

to histamine, but the difference was not statistically significant. The infusion rate for a maximum response was 0.0032 μmol per kg per min. In conscious Heidenhain pouch dogs, impromidine had ~22 times the potency of histamine (Fig. 2). The maximum response to impromidine was significantly greater than that obtained to histamine ($P = 0.05$–0.025). Maximal secretion was obtained at an infusion rate of 0.0008 μmol per kg per min (compare with histamine 0.022 μmol per kg per min). In all species studied the secretory response to impromidine was inhibited by histamine H_2-receptor antagonists. In the Heidenhain pouch dog, an infusion of cimetidine at 10 μmol per kg per h produced a mean inhibition of 86% ($n = 6$) of maximal acid secretion; this is comparable with that obtained against histamine for the same dose of cimetidine.

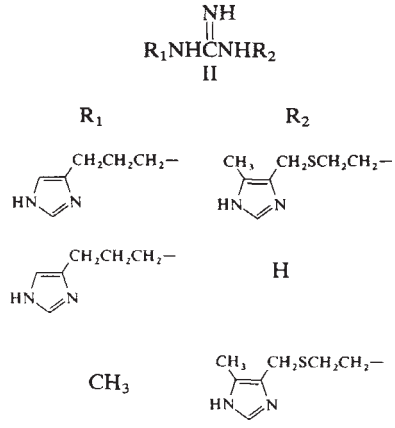
Blood pressure and gastric acid secretion were measured simultaneously in the anaesthetised rat and cat preparations and also in the acutely fistulated anaesthetised dog. The results shown in Table 2 indicate that at doses maximal for the stimulation of acid secretion, impromidine, in contrast to histamine, produced only small changes in mean blood pressure and in pulse pressure. Impromidine has been shown to be a specific agonist at histamine H_2 receptors. The lack of marked effects on the cardiovascular system at a dose level maximal for acid secretion suggests that impromidine may be useful as a diagnostic agent in man for the estimation of the maximal secretory capacity of the stomach in patients with gastrointestinal disease. Preliminary studies are reported elsewhere[10,11].

Subacute repeated dose studies have been carried out in the rat and dog for up to 14 d at dose levels >10 times the dose required for maximal stimulation of gastric acid secretion. No toxic effects which might limit the use of impromidine were apparent from these studies (G. B. Leslie, unpublished observations). Full details will be reported elsewhere.

The most interesting feature of the pharmacology of impromidine is its high potency compared to histamine. From the pharmacological analysis it seems that this increased potency is the result of an increased affinity of the compound for H_2 receptors rather than an increase in efficacy (using Stephenson's nomenclature[12]). This is obvious for the uterus preparation where impromidine is not a full agonist, although more potent than histamine; it would seem that in comparison with histamine many more receptors have to be occupied by impromidine to achieve a response. The high affinity of impromidine relative to histamine should make this a useful agent for binding studies at H_2 receptors.

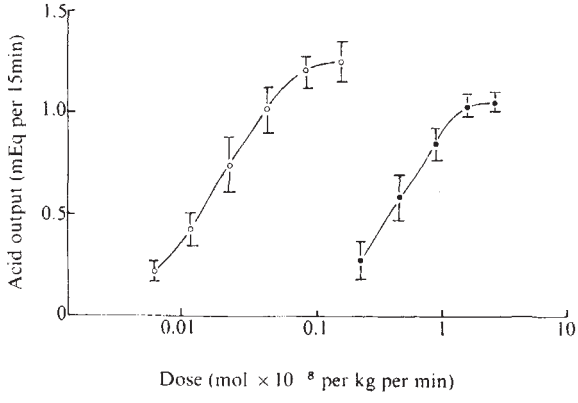
The chemical structure poses interesting questions in relation to the pharmacological activity. Impromidine is a guanidine derivative having two different imidazolylalkyl substituents (Table 3, R_1 and R_2 in II), and it is instructive to compare it with analogous guanidines (Table 3, III and IV). The guanidine III (SK&F 91486) is a weak partial agonist at H_2 receptors[13] and must have a low efficacy relative to histamine. Adding the substituent R_2 to structure III gives impromidine and results in a full agonist on the atrium and in a 1,000-fold increase in potency. The group R_2 is, however, a structural feature of H_2-receptor antagonists (for example, metiamide[14,15] and cimetidine[8]) and, furthermore, the analogous guanidine IV (SK&F 92408) is an antagonist showing no agonist activity[16], suggesting that R_2 is associated with affinity but not with efficacy. Thus, a chemical group which contributes affinity in antagonist structures apparently converts a weak partial agonist (Table 3, III) into an agent (impromidine) which has increased efficacy as well as increased affinity. This result has important implications in considering drug–receptor interaction.

Table 3 Chemical structures of impromidine (SK&F 92676) and related guanidines



SK&F 91486 has 4% of the potency of histamine on the guinea pig isolated atrium but achieves only 70% of the maximum response relative to histamine; it competitively antagonises histamine ($pA_2 = 4.65$)[13]. These reported[13] results were obtained in the absence of propranolol; subsequent studies have shown that in the presence of propranolol, SK&F 91486 only attains 30.8±3.0% (mean±s.e.m.) of histamine's maximum response. On the rat isolated uterus (in the presence of propranolol) it has 6.5% of the potency of histamine and achieves only 27% of the maximal response[13]. SK&F 92408 competitively antagonises histamine[16], pA_2 4.80 (atrium), 5.26 (uterus).

Fig. 2 Dog gastric secretion; dose–response curves to intravenous histamine (●) and SK&F 92676 (○) in the conscious Heidenhain pouch dog. Each point is the mean of five experiments±s.e.m. Dose–response curves to intravenous infusion of the agonist were constructed using a stepwise doubling increment in the dose, each dose level being infused for 1 h and the mean of the acid output in the last two 15-min samples being taken to calculate the dose–response curve.



We thank D. W. Hills for his contribution to the original synthesis of impromidine, and Dr D. H. Jenkinson for discussion.

G. J. DURANT
W. A. M. DUNCAN
C. R. GANELLIN
M. E. PARSONS
R. C. BLAKEMORE
A. C. RASMUSSEN

The Research Institute,
Smith Kline & French Laboratories Limited,
Welwyn Garden City, Hertfordshire, UK

Received 13 July; accepted 9 October 1978.

1. Black, J. W., Duncan, W. A. M., Durant, G. J., Ganellin, C. R. & Parsons, M. E. *Nature* **236**, 385–390 (1972).
2. Bertaccini, G., Impicciatore, M., Vitali, T. & Plazzi, V. *Farmaco* **27**, 680–682 (1972).
3. Bertaccini, G., Impicciatore, M. & Vitali, T. *Farmaco* **31**, 934–938 (1976).
4. Impicciatore, M., Mossini, F., Plazzi, V., Chiavarini, M. & Bertaccini, G. *Atteneo Parmense, Acta Bio-Med.* **45**, 145–148 (1976).
5. Durant, G. J., Emmett, J. C., Ganellin, C. R., Roe, A. M. & Slater, R. A. *J. med. Chem.* **19**, 923–928 (1976).
6. Grossman, M. I., Robertson, C. & Rosiere, C. E. *J. Pharm. exp. Ther.* **104**, 277–283 (1952).

7. Parsons, M. E., Owen, D. A. A., Ganellin, C. R. & Durant, G. J. *Agents Actions* **7**, 31–37 (1977).
8. Brimblecombe, R. W. *et al. J. Int. med. Res.* **3**, 86–92 (1975).
9. Parsons, M. E. thesis, Univ. London (1969).
10. Hunt, R. H., Billings, J. A., Mills, J. G., Burland, W. L. & Milton-Thompson, G. J. *Gut* (in the press).
11. Halter, F., Bangerter, U., Schlup, M., Liechti, M. & Hacki, W. H. *Gut* (in the press).
12. Stephenson, R. P. *Br. J. Pharmac.* **11**, 379–393 (1956).
13. Parsons, M. E., Blakemore, R. C., Durant, G. J., Ganellin, C. R. & Rasmussen, A. C. *Agents Actions* **5**, 464 (1975).
14. Black, J. W. *et al. Agents Actions* **3**, 133–137 (1973).
15. Black, J. W., Durant, G. J., Emmett, J. C. & Ganellin, C. R. *Nature* **248**, 65–67 (1974).
16. Durant, G. J. *et al., J. med. Chem.* **20**, 901–906 (1977).

A new dopaminergic prodrug

THE clinical treatment of parkinsonism with the dopamine (DA) precursor L-dopa has certain drawbacks[1] and there are relatively few drugs which directly stimulate DA receptors (although there are many which block them, such as the neuroleptics). There is thus great industrial, academic and clinical interest in the development of new DA-receptor agonists[1]. Recently much interest has been given to 2-amino-6, 7-dihydroxytetrahydronaphthalene (ADTN)(Fig. 1), which is known from various *in vitro* and *in vivo* studies[2–10] to be a potent, selective and long-acting DA agonist. Two drawbacks to its more widespread study, however, are that the original synthesis was not easy, consisting of eight steps[11], and like DA, it does not readily pass the blood–brain barrier, so that it must be administered to animals intraventricularly. Although *in vivo* ADTN is not the most potent member of its group of 2-aminotetralin derivatives[12], it has the advantage of being structurally very similar to DA, but metabolically more stable. We have developed an improved, simplified synthesis of ADTN consisting of four steps[13] and we now describe how we have effected the passage of the drug into the brain. In addition to their practical and clinical implications, our results throw new light on the *in vivo* action of this drug.

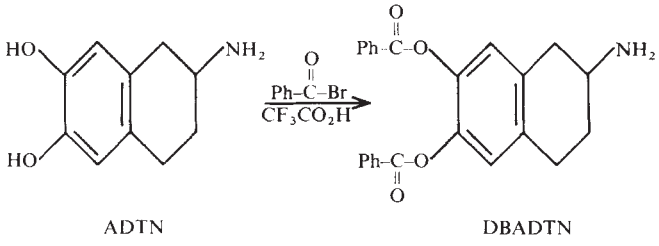


Fig. 1 The molecular weight of ADTN is 179 and that of DBADTN 387; thus, a dose of 50 mg per kg DBADTN is equivalent to 23 mg per kg ADTN.

As lipid-soluble drugs rapidly pass the blood–brain barrier, we synthesised a 'prodrug' from ADTN. A prodrug is a derivative that has certain desirable properties, for example, enhanced lipid solubility, which the parent molecule does not possess, and that can be converted back (by enzymes) into the active species within the body. The synthesis consisted of making a diester derivative of the catechol function (Fig. 1) which reduces the polarity of the parent molecule and makes it more lipid soluble. It was anticipated that after entering the brain the diester-prodrug would be converted back to ADTN by enzymatic hydrolysis. The dibenzoyl ester (DBADTN) was chosen as it was predicted on theoretical grounds that its rate of hydrolysis would be slower than that of the simple aliphatic esters[14], an important point if sufficient prodrug is to survive hydrolysis outside the brain, to pass into this organ. The diester was readily prepared in a yield of 74% by treating ADTN in trifluoroacetic acid with benzoyl bromide at room temperature (Fig. 1). The presence of trifluoroacetic acid ensures that only the catechol function is esterified and that the amino group is not affected. Analytical and spectral data were consistent with the proposed structure, and details will be published elsewhere.

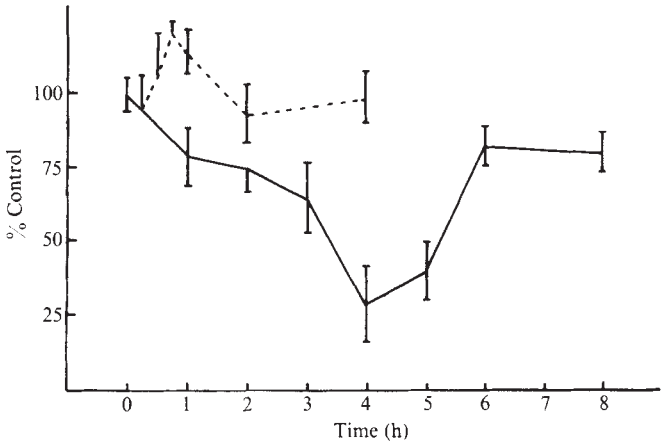


Fig. 2 Effect of DBADTN on HVA and DA concentrations in rat corpus striatum. Female Wistar rats (150–200 g) received 50 mg per kg DBADTN i.p. in a vehicle consisting of 30% ethanol, 30% water and 40% polyethylene glycol 400. HVA and DA were determined in rat corpus striatum using previously described methods[19,30]. Control values (mean ± s.e.m. (number of determinations)) for HVA were 0.52 ± 0.03 (27) μg per g and for DA, 5.68 ± 0.29 (5) μg per g wet weight. A statistical analysis consisting of an analysis of variance followed by the Newman–Keuls test[31] on the absolute values of HVA and DA showed that the decrease in HVA concentration at 3, 4 and 5 h was significantly different from control values at the 5%, 1% and 1% level, respectively. None of the changes in DA concentration were significant at the 5% level. Each point is the mean ± s.e.m. of 7–8 determinations for HVA and 4–8 for DA. – – –, DA; ——, HVA.

The idea that the prodrug would enter the brain and be hydrolysed to ADTN was tested by injecting 100 mg per kg DBADTN intraperitoneally (i.p.) into a female Wistar rat, killing the animal 45 min later, and rapidly dissecting the corpus striatum from the rest of the brain. The ADTN was isolated and derivatised with pentafluoropropionic anhydride using a modification of the method of Wilk and Stanley[15] for gas chromatography–mass spectrometry mass fragmentographic analysis. By monitoring the peaks at *m/e* 307 and 454 from the acylated derivative of ADTN, it was unequivocally shown that ADTN was present in the corpus striatum and the rest of the brain. DA agonists are known to decrease DA turnover in the corpus striatum as reflected by a lowering of the concentration of homovanillic acid (HVA)[16–18], one of the main metabolites.

The effect of 50 mg per kg DBADTN (equivalent to 23 mg per kg ADTN) on HVA levels in the corpus striatum of rats was studied over 8 h (Fig. 2). A 70% decrease of control levels was found after 4 h but by 6 h the HVA concentrations were almost normal. This slow onset of the maximum effect, in comparison with the rapid effect of apomorphine[19], is probably a result of the delayed concentration rise of ADTN in the brain brought about by enzymatic hydrolysis from the prodrug. Preliminary experiments indicate that lower doses of the prodrug also produce decreases in striatal HVA levels in a dose-related manner. As α-methyldopa is known to produce a very similar effect on HVA levels after the same time period[20] as well as lowering endogenous striatal DA levels, and as this may be brought about by its conversion to α-methyldopamine (which can be viewed as a ring-opened lower homologue of ADTN), the effect of DBADTN on striatal DA levels was also examined (Fig. 2). It was found however, that there was no statistically significant depletion of DA over a 4-h period. To ensure that the effect on

0028–0836/78/0276—0405$01.00
© Macmillan Journals Ltd 1978

HVA levels was due to ADTN produced in the brain and not to entry of ADTN produced by enzymatic hydrolysis outside the brain, 30 mg per kg of ADTN itself was injected i.p. into rats; however, after 4 h there was no significant lowering of HVA levels (92 ± 3% ($n = 8$) of controls) in the striatum.

It is known that the increase in DA turnover, as reflected in elevated HVA levels, which is caused by DA-receptor antagonists such as the neuroleptics, can be attenuated by DA agonists[17,21]. We were also able to demonstrate this with DBADTN in a dose-dependent manner; thus, the effects of 0.50, 0.25 and 0.05 mg per kg haloperidol could all be attenuated by administration of 50 mg per kg DBADTN (Fig. 3).

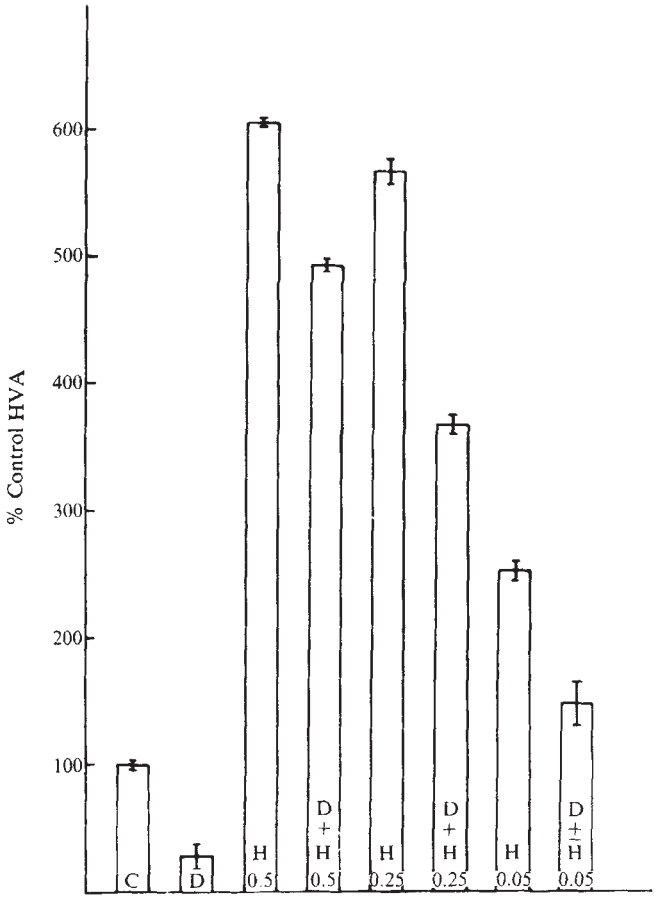


Fig. 3 Effects of DBADTN and various doses of haloperidol, alone and in combination, on HVA levels in rat corpus striatum. Female Wistar rats (150–200 g) received 50 mg per kg DBADTN (D) i.p. in the standard vehicle and 2.5 h after 0.50, 0.25 or 0.05 mg per kg haloperidol (H) i.p., the animals were then killed 1.5 h after the haloperidol injection. The HVA concentration in the corpus striatum was determined as previously described[19], control values (mean ± s.e.m. (number of determinations)) were 0.36 ± 0.01 (20) μg per g wet weight. Statistical analysis using Students t test on the HVA values in the experimental groups, haloperidol (H) against haloperidol (H) + DBADTN (D) showed that for the haloperidol doses of 0.5, 0.25, and 0.05 mg per kg the levels of significance were $P < 0.01$ and < 0.01 and < 0.01, respectively. Each group is the mean ± s.e.m. for 7–8 determinations.

These results agree with the inhibition of the behavioural effects of ADTN by 0.5 mg per kg haloperidol[6]. It has now been shown that there are at least three DA receptors in the CNS, a postsynaptic and two 'presynaptic' receptors, one of which is probably located on the presynaptic nerve terminal, with the other near the DA cell body[22]. We have already presented *in vitro* evidence for an action of ADTN on the latter of these presynaptic DA receptors, that is, in the substantia nigra[23]. It was shown that in homogenates of this brain area, ADTN was as potent as DA in stimulating cyclic AMP production. It was

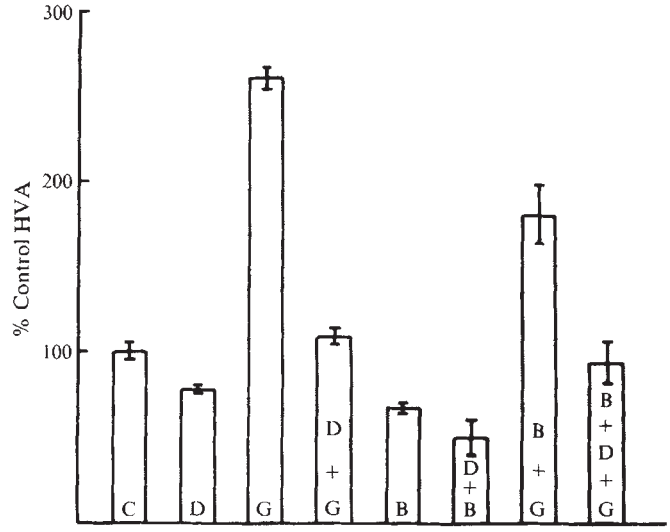


Fig. 4 Effect of DBADTN pretreatment on the GBL-induced rise in HVA levels in rat corpus striatum. Female Wistar rats (150–200 g) received 25 mg per kg benztropine[28] (B) i.p. and 10 min later 50 mg per kg DBADTN (D). 30 min after injection of D they were injected with 750 mg per kg γ-butyrolactone, GBL (G), and 90 min after this the animals were killed and HVA concentrations in the corpus striatum determined[19]. Control values were 0.54 ± 0.03 (14) μg per g wet weight (mean ± s.e.m.(number of determinations)). Statistical analysis (analysis of variance followed by the Newman–Keuls test[31]) on the HVA values showed that the differences between control (C) compared with G, and G compared with D + G were both significant at the 1% level. There was no significant difference at the 5% level between the groups D + G compared with B + D × G. Each group is the mean ± s.e.m. of 7–16 determinations.

therefore of interest to examine the effect of DBADTN (and hence ADTN) on the terminal presynaptic DA receptor using the γ-butyrolactone (GBL) model[22]. Administration of GBL to rats results in an inhibition of impulse flow in the dopaminergic presynaptic axons which causes a decrease in DA release coupled with a marked increase in its biosynthesis; this leads to a large rise in DA[22] and HVA[24] levels. The inhibition of these elevations by DA agonists is considered to be due to stimulation of presynaptic DA receptors on the axon terminal[22,24,25]. As shown in Fig. 4 the rise in HVA levels caused by GBL could be attenuated by pretreatment with DBADTN (compare G with D + G). This is compelling evidence for an action of ADTN on the presynaptic DA terminal[24]. Attempts to inhibit this effect of ADTN on HVA levels by blocking its uptake with benztropine[26] were without success, suggesting possibly that *in vivo* interaction with membrane presynaptic receptors is of more significance than is cytoplasmic accumulation of ADTN. *In vitro*, however, it has been shown[27] that benztropine does affect the ability of ADTN to interfere with DA turnover. Benztropine itself produced a lowering of HVA levels that was not statistically significant at the 5% level; Braestrup has also reported that it has no significant effect[28].

We have demonstrated that it is possible to prepare a successful prodrug from a molecule that does not pass the blood–brain barrier. It should be stressed that the situation here is quite different from that occurring with the recently described esters of apomorphine[29] in that apomorphine itself can pass the blood–brain barrier and a reasonable amount of the free drug produced in the body by hydrolysis will probably pass into the brain. Second, we have shown that in addition to being able to interact with postsynaptic and nigral DA receptors, ADTN can also interact *in vivo* with DA receptors which are thought to be located on DA nerve terminals. We are now attempting to determine the concentration of ADTN in the corpus striatum and other brain areas and also to examine the effects of the

prodrug on animal behaviour. The latter results may indicate whether or not the prodrug has any possible clinical potential.

A. S. HORN
D. DE KASTE
D. DIJKSTRA
H. ROLLEMA
M. FEENSTRA
B. H. C. WESTERINK
C. GROL
A. WESTERBRINK

Department of Pharmacy,
University of Groningen,
9713 AW, Groningen, The Netherlands

Received 31 May; accepted 26 September 1978.

1. Hornykiewicz, O. *Biochem. Pharmac.* **24,** 1061–1065 (1975)
2. Woodruff, G. N., Elkhawad, A. O., Crossman, A. R. & Walker, R. J. *J. Pharm. Pharmac.* **26,** 740–741 (1974).
3. Miller, R. J., Horn, A. S., Iversen, L. L. & Pinder, R. *Nature* **250,** 238–241 (1974).
4. Woodruff, G. N., Elkhawad, A. O. & Pinder, R. M. *Eur. J. Pharmac.* **25,** 80–86 (1974).
5. Horn, A. S. *J. Pharm. Pharmac.* **26,** 740–741 (1974).
6. Elkhawad, A. O. & Woodruff, G. N. *Br. J. Pharmac.* **54,** 107–114 (1975).
7. Costall, B., Naylor, R. J. & Pinder, R. M. *Eur. J. Pharmac.* **31,** 94–109 (1975).
8. Cramly, H. J., Pinder, R. M., Banks Hinstraw, W. & Goldberg, L. I. *Nature* **259,** 584–587 (1976).
9. Burt, D. M., Creese, I. & Snyder, S. H. *Molec. Pharmac.* **12,** 800–812 (1976).
10. Woodruff, G. N., Kelly, P. H. & Elkhawad, A. O. *Psychopharmacology* **47,** 195–198 (1976).
11. Thrift, R. I. *J. chem. Soc. C* 288–293 (1967).
12. Costall, B., Naylor, R. J., Cannon, J. G. & Lee, T. *Eur. J. Pharmac.* **41,** 307–319 (1977).
13. Horn, A. S., Grol. C. J., Dijkstra, D. & Mulder, A. H. *J. med. Chem.* **21,** 825–828 (1978).
14. Gould, E. S. in *Mechanism and Structure in Organic Chemistry,* 318 (Holt, Rinehart & Winston, London, 1969).
15. Wilk, S. & Stanley, M. *Eur. J. Pharmac.* **41,** 65–72 (1977).
16. Roos, B. E. *J. Pharm. Pharmac.* **21,** 263–264 (1969).
17. Jori, A., Cecchetti, G., Dolfini, E., Monti, E. & Garattini, S. *Eur. J. Pharmac.* **27,** 245–248 (1974).
18. Dolphin, A. C., Jenner, P., Sawaya, M. C. B., Marsden, C. D. & Testa, B. *J. Pharm. Pharmac.* **29,** 727–734 (1977).
19. Westerink, B. H. C. & Korf, J. *Eur. J. Pharmac.* **40,** 131–136 (1976).
20. Uretsky, N. J. *J. Pharmac. exp. Ther.* **189,** 359–369 (1974).
21. Lahti, R. A., McAllister, B. & Wozniak, J. *Life Sci.* **11,** 605–613 (1972).
22. Walters, J. R. & Roth, R. H. *Naunyn-Schmiedebergs Archs Pharmak.* **296,** 5–14 (1976).
23. Phillipson, O. T. & Horn, A. S. *Nature* **26,** 418–420 (1976).
24. Martres, M. P., Costentin, J., Bandry, M., Marcais, H., Protais, P. & Schwartz, J. C. *Brain Res.* **136,** 319–337 (1977).
25. Walters, J. R. & Roth, R. H. *J. Pharmac. exp. Ther.* **191,** 82–91 (1974).
26. Davis, A., Roberts, P. J. & Woodruff, G. N. *Br. J. Pharmac.* **63,** 183–190 (1978).
27. Iversen, L. L., Rogawsky, M. A. & Miller, R. J. *Molec. Pharmac.* **12,** 251–262 (1976).
28. Braestrup, C. *J. Pharm. Pharmac.* **29,** 463–470 (1977).
29. Baldessarini, R. J., Kula, N. S., Walton, K. G. & Borgman, R. J. *Psychopharmacology* **53,** 45–53 (1977).
30. Westerink, B. H. C. & Korf, J. *J. Neurochem.* **29,** 697–706 (1977).
31. Zivin, J. A. & Bartko, J. J. *Life Sci.* **18,** 15–26 (1975).

Alterations in β-adrenergic receptor binding during ethanol withdrawal

CHARACTERISTIC features of ethanol withdrawal, including tremor, agitation, seizures, tachycardia and increased blood pressure, frequently develop in chronic alcoholics 6–60 h after the discontinuation of long-term heavy ethanol consumption[1–4]. Similar symptoms are seen in hyperadrenergic states such as anxiety and thyrotoxicosis[5,6]. There is evidence to suggest that augmented sensitivity to catecholamines in thyrotoxicosis is due to the development of β-adrenergic receptor supersensitivity in cardiac and skeletal muscle cell surfaces[7–10]. Propranolol, a β-adrenergic blocker, has been shown to control some of the clinical signs of ethanol withdrawal, such as arrhythmia, hypertension and tremor[11–13]. In addition, propranolol significantly decreases urinary excretion of noradrenaline during ethanol withdrawal[14]. These findings have led to the suggestion that peripheral manifestations of increased adrenergic activity in ethanol withdrawal are secondary to increased sympathetic neural outflow and that this increase in central catecholaminergic outflow may be the result of increased β-adrenergic receptor sensitivity in the alcohol withdrawal state in man[14]. We report here attempts to determine whether increased central and peripheral β-adrenergic receptor binding could be demonstrated in rats during ethanol withdrawal.

Table 1 Effect of chronic ethanol treatment on specific binding of ^{3}H-DHA to β-adrenergic receptors in brain and heart particulate fractions

Time after termination of ethanol treatment (h)	Brain		Heart	
	Control	Ethanol-treated	Control	Ethanol-treated
0	37.7±0.57	29.1±0.96*	26.2±1.29	14.9±0.59*
24	33.5±2.81	36.1±3.39	28.2±1.37	32.7±1.05
48	26.9±2.36	47.2±4.00*	27.1±0.19	40.5±2.61*
72	31.6±1.66	52.8±4.19*	28.8±2.16	67.9±3.25*

Rats were pair-fed a nutritionally complete liquid diet containing either 6.5% (v/v) ethanol or an equicaloric amount of sucrose for 60 d by a previously described method[15]. Ethanol provided 36% of the total calories for the chronically treated rats (daily intake approximately 12–14 g per kg). The purity of ^{3}H-dihydroalprenolol (DHA) (NEN) was tested by thin-layer chromatography[16]. The procedures for the preparation of brain and heart particulate fractions and measurement of specific binding of ^{3}H-DHA to β-adrenergic receptors have been described previously[17–19]. The results are shown as fmol of DHA bound per mg protein and are the means (±s.e.m.) of 5 or 6 determinations (each in triplicate) from 5 or 6 rats in each group. The concentration of ^{3}H-DHA was 2 nM in each binding assay.

* Significantly different from control ($P<0.05$), by Student's *t*-test.

Specific binding of ^{3}H-dihydroalprenolol (DHA) to brain and heart particulate fractions obtained from sucrose- and ethanol-fed rats is shown in Table 1. The specific binding of ^{3}H-DHA in heart and brain particulate fractions decreased by 43% and 23% respectively in the ethanol-fed group as compared with the sucrose-fed control rats at zero hour of termination of chronic ethanol treatment. The high concentration of ethanol in rat blood at this time suggests that the decrease in ^{3}H-DHA binding might be due to interference by ethanol with the drug–receptor interaction. To test this possibility, specific binding of ^{3}H-DHA to heart and brain particulate fractions was measured in the absence or presence of ethanol in the binding assay medium. Ethanol (2 mg ml^{-1}) did not decrease binding of ^{3}H-DHA to heart or brain particulate fractions. In addition, 10 rats were divided into two groups; group I being treated intraperitoneally with ethanol, 2.5 g per kg body weight, administered as a 12.5% (w/v) solution in normal saline, and group II, which served as the control, receiving an equal volume of normal saline. Rats were killed 2 h after drug administration, and brain and heart particulate fractions were prepared as described in Table 1. Specific binding of ^{3}H-DHA to these preparations obtained from control and ethanol-treated rats were the same. These observations suggest that the decrease in specific binding of ^{3}H-DHA to the heart and brain particulate fractions immediately after termination of chronic ethanol treatment is not due to interference by ethanol with the drug–receptor interaction.

One day after discontinuation of chronic ethanol treatment, the specific binding of ^{3}H-DHA to heart and brain particulate fractions was similar in sucrose- and ethanol-fed rats. On the other hand, 48 h after cessation of ethanol administration, the specific binding of ^{3}H-DHA increased by 75% in brain and 49% in heart as compared with control preparations. Seventy-two hours after withdrawal from chronic alcohol administration the specific binding of ^{3}H-DHA was enhanced by 67% in brain and 136% in heart.

The increased binding of ^{3}H-DHA in rat brain particulate fractions during the withdrawal state following chronic ethanol treatment suggests changes in either the number or the apparent affinity of β-adrenergic receptors for ^{3}H-DHA. Therefore, the equilibrium dissociation constants and maximal number of binding sites for the β-adrenergic receptor were determined graphically (Fig. 1). Enhancement of the binding of ^{3}H-DHA to β-adrenergic receptors in rat brain particulate fractions 48 and

© Macmillan Journals Ltd 1978

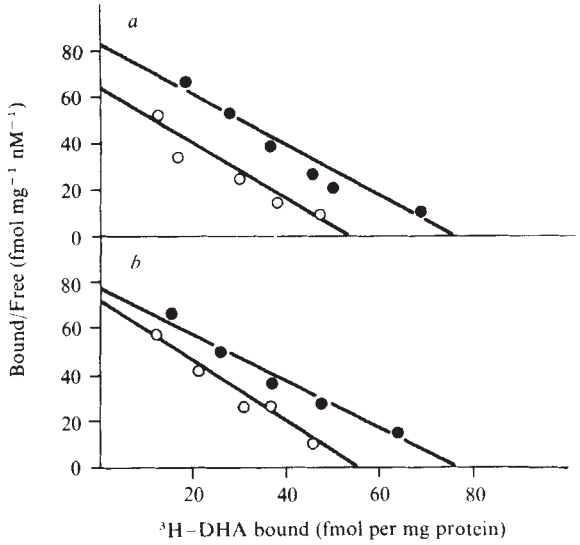


Fig. 1 Scatchard analysis of specific binding of ^{3}H-DHA in brain particulate fractions obtained from sucrose (○) and ethanol (●) fed rats at 48 h (*a*) and 72 h (*b*) after cessation of ethanol administration. The concentrations of ^{3}H-DHA used varied from 0.125 to 4 nM. Slopes and intercepts of the lines were determined by the method of least squares. The correlation coefficient, $r \geq 0.82$.

72 h after cessation of chronic ethanol intake appears to be due to an increase in the number of binding sites (Table 2).

Propranolol, a β-adrenergic receptor antagonist, has been shown to be of value in the management of withdrawal reactions of chronic alcoholism including augmented urinary excretion of noradrenaline[11-14]. These observations have led to the hypothesis that the peripheral effects of ethanol withdrawal may be a consequence of increased central β-adrenergic receptor sensitivity. The results described in the present communication provide experimental evidence in support of this hypothesis. In addition, present observations show that increased β-adrenergic receptor binding may occur both in the central nervous system and in peripheral tissues during ethanol withdrawal. Thus, the therapeutic efficacy of propranolol during ethanol withdrawal reactions may be mediated in part by its antagonism to the effects of increased central catecholaminergic outflow and in part by directly blocking hypersensitive β-adrenergic receptors in peripheral tissues.

The ethanol withdrawal state has been divided into two major components[1-4]. In mild reactions, the chief symptoms are sleeplessness and irritability, which appear within a few hours of withdrawal and disappear after 48 h. In severe reactions these symptoms are followed by tremulousness, hallucinosis, seizures, and arrhythmias which are most evident 2–3 d after ethanol withdrawal[1-4]. This suggests that the increase in β-adrenergic receptor sensitivity may contribute to the withdrawal reaction following the interruption of chronic ethanol intake.

Attenuated as well as increased accumulations of noradrenaline-stimulated cerebral cortical cyclic AMP have been reported with chronic ethanol treatment in rats and mice[20-22]. As noradrenaline may increase formation of cyclic AMP by interacting with either α-adrenergic or β-adrenergic receptors in brain, the results of French, Israel and their associates[20-22] fail to provide any conclusive evidence for a specific alteration of sensitivity of α-adrenergic or β-adrenergic receptors in chronic alcoholism.

The molecular mechanism for the increase in β-adrenergic receptor binding in rat heart and brain particulate fractions during the ethanol withdrawal state is not known. A variety of psychotropic drugs, such as antipsychotics[23], antidepressants[17] and amphetamine[19], have been shown to alter the sensitivity of dopaminergic and β-adrenergic receptors. Amphetamine and desipramine are known to affect noradrenaline transport at the catecholaminergic nerve endings[24,25]. The Na^+-dependent noradrenaline transport appears to be mediated through the activity of $(Na^+ + K^+)$-ATPase[26,27]. A marked activation in the number of $(Na^+ + K^+)$-ATPase molecules in several areas of brain and some peripheral organs has been observed during ethanol withdrawal in cats and rats[28-31]. It is suggested that this activation of $(Na^+ + K^+)$-ATPase may contribute to increased β-adrenergic receptor sensitivity in rat heart and brain following interruption of chronic ethanol intake.

Table 2 Effect of chronic ethanol treatment on the equilibrium dissociation constant (K_D) and maximal number of binding sites (B_{max}) of β-adrenergic receptors in rat brain

Time after termination of ethanol treatment (h)	Control B_{max} (fmol per mg protein)	Control K_D (nm)	Ethanol-treated B_{max} (fmol per mg protein)	Ethanol-treated K_D (nM)
0	62±6	0.71±0.06	45+7*	0.81±0.07
48	53±8	0.85±0.07	76±11*	0.91±0.06
72	55±9	0.77±0.05	76+9*	0.98±0.05

The specific binding of 5 concentrations of ^{3}H-DHA (0.125–4 nM) to β-adrenergic receptors in rat brain particulate fractions was assayed as described in Table 1. The equilibrium dissociation constants and maximal number of binding sites were estimated by Scatchard analysis, as described before[17-19]. Results shown are the mean (±s.e.m.) of 5 determinations (each in triplicate) from 5 rats for each of the 5 concentrations of ^{3}H-DHA used.

* Significantly different from control ($P < 0.05$), by Student's *t*-test.

The studies presented here suggest that increased β-adrenergic receptor binding in rat brain and peripheral tissues may contribute to the hyperadrenergic state of the withdrawal reaction of chronic alcoholism. These observations provide a rationale for the use of β-adrenergic antagonists in the management of the ethanol withdrawal reaction.

These studies were supported in part by grants from the NIHL and AHA, and a post-doctoral fellowship (to V.K.S.) from the Genessee Valley Heart Association. We thank Dr Ann Young for stimulating discussions.

SHAILESH P. BANERJEE
VIRENDRA K. SHARMA
Department of Pharmacology and Toxicology, University of Rochester School of Medicine and Dentistry, Rochester, New York 14642

JATINDER M. KHANNA
Department of Pharmacology, University of Toronto, and Addiction Research Foundation of Ontario, Toronto, Canada

Received 7 July; accepted 29 September 1978.

1. Victor, M. *Psychosom. Med.* **28,** 636–650 (1966).
2. Victor, M. *Postgrad. Med.* **47,** 68–72 (1970).
3. Gross, M. M., Lewis, E. & Hastey, J. in *Clinical Pathology* Vol. 3 (eds Kissin, B. & Begleiter, H.) 191–263 (Plenum, New York, 1974).
4. Sellers, E. M. & Kalant, H. *New Engl. J. Med.* **294,** 757–762 (1976).
5. Marsden, C. D., Gimlette, T. M., McAllister, R. G., Owen, D. A. L. & Miller, T. N. *Acta Endocrinol.* **57,** 353–362 (1968).
6. Marsden, C. D., Meadows, J. C. & Lange, G. W. *J. Neurol. Neurosurg. Psychiat., Lond.* **33,** 776–782 (1970).
7. Banerjee, S. P. & Kung, L. S. *Eur. J. Pharmac.* **43,** 207–208 (1977).
8. Sharma, V. K. & Banerjee, S. P. *Biochim. biophys. Acta* **539,** 538–542 (1978).
9. Ciaraldi, T. & Marinetti, G. V. *Biochem. biophys. Res. Commun.* **74,** 984–991 (1977).
10. Williams, L. T., Lefkowitz, R. J., Watanabe, A. M., Hathaway, D. R. & Besch, H. R., Jr *J. biol. Chem.* **252,** 2787–2789 (1977).
11. Carlsson, C. & Johannson, T. *Br. J. Psychiat.* **119,** 605–606 (1971).
12. Zilm, D. H., Sellers, E. M., Macleod, S. M. & Degani, N. C. *Ant. Int. Med.* **83,** 234–236 (1974).
13. Sellers, E. M., Zilm, D. H., Degani, N. C. & MacLeod, S. M. *Clin. Res.* **23,** 610A (1975).
14. Sellers, E. M., Degani, N. C., Zilm, D. H. & MacLeod, S. M. *Lancet* **i,** 94–95 (1976).
15. Khanna, J. M., Kalant, H. & Bustos, G. *Can. J. physiol. Pharmac.* **45,** 777–785 (1967).
16. Bylund, D. R. & Snyder, S. H. *Molec. Pharmac.* **12,** 568–580 (1976).
17. Banerjee, S. P., Kung, L. S., Riggi, S. J. & Chanda, S. K. *Nature* **268,** 455–456 (1977).
18. Sharma, V. K. & Banerjee, S. P. *J. biol. Chem.* **252,** 7444–7446 (1977).
19. Banerjee, S. P., Sharma, V. K., Kung, L. S. & Chanda, S. K. *Nature* **271,** 380–381 (1978).
20. Israel, M. A., Kimura, H. & Kuriyama, K. *Experientia* **28,** 1322–1323 (1972).
21. French, S. W., Reid, P. E., Palmer, D. S., Narod, M. E. & Ramey, C. W. *Res. Commun. chem. pathol. Pharmac.* **6,** 651–662 (1973).
22. French, S. W., Palmer, D. S., Narod, M. E., Reid, P. E. & Remey, C. W., *J. Pharmac. exp. Ther.* **194,** 319–326 (1975).

23. Burt, D. R., Creese, I. & Snyder, S. H. *Science* **196**, 326–327 (1977).
24. Groves, P. M. & Rebec, G. V. *Ann. Rev. Psychol.* **27**, 91–128 (1976).
25. Sulser, F., Vetulani, J. & Mobley, P. L. *Biochem. Pharmac.* **27**, 257–262 (1978).
26. Iversen, L. L. In *Handbook of Psychopharmacology*, Vol III (eds Iversen, L. L., Iversen, S. D. & Snyder, S. H.) 381–442 (Plenum, New York).
27. Sharma, V. K. & Banerjee, J. P. *Eur. J. Pharmac.* **41**, 417–429 (1977).
28. Knox, W. H., Perrin, R. G. & Sen, A. K. *J. Neurochem.* **19**, 2881–2884 (1972).
29. Sharma, V. K., Nagraj, M. G. & Banerjee, S. P. *Brain Res.* **129**, 183–186 (1977).
30. Banerjee, S. P. & Sharma, V. K. *Br. J. Pharmac.* **62**, 475–579 (1978).
31. Rangaraj, N. & Kalant, H. *Biochem. Pharmac.* (in the press).

Epidermal growth factor stimulates phosphorylation in membrane preparations *in vitro*

EPIDERMAL GROWTH FACTOR (EGF) forms a complex with plasma membrane receptors in intact cells that initiates a series of biochemical events resulting in increased cell growth *in vivo* and *in vitro*[1]. The interaction of EGF with membrane receptors has been demonstrated in crude membrane preparations[2], but no biochemical alteration of the membrane resulting from hormone binding has been detected. To clarify the molecular mechanisms regulating cell proliferation, specific biochemical reactions initiated by mitogens such as EGF need to be investigated in cell-free systems. As the human epidermoid carcinoma cell line A-431 has an extraordinarily high concentration of EGF receptors[3,4] ($2–3\times10^6$ receptors per cell), we have used a crude membrane preparation from these cells to look for an EGF-dependent alteration of membrane structure and/or function. We report here that (1) membranes may be prepared from A-431 cells that retain the ability to bind ^{125}I-labelled EGF in a specific manner, and (2) the binding of EGF to these membranes *in vitro* results in a marked stimulation of the phosphorylation of endogenous proteins in the presence of [γ-^{32}P]ATP.

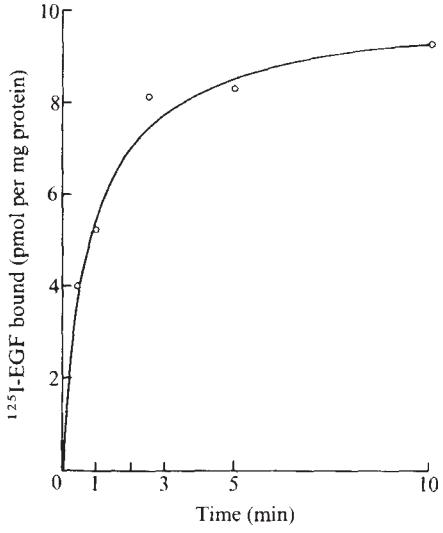


Fig. 1 Binding of ^{125}I-EGF at 0 °C to A-431 membranes. Membranes were prepared by the procedure described by Thom *et al.*[5] from A-431 tumour cells grown in Falcon culture dishes containing Dulbecco's modified Eagle's medium supplemented with 10% fetal calf serum and gentamycin. ^{125}I-EGF was prepared as reported by Carpenter and Cohen[6]. The binding reaction was carried out in a 200-μl reaction volume containing 20 mM HEPES buffer, *p*H 7.4, 0.1% bovine serum albumin (BSA), 12 μg of A-431 membrane protein (measured by the Bradford procedure[7] using γ-globulin as a standard), and ^{125}I-EGF (3.3×10^{-9}M). Nonspecific binding was measured in co-lateral reaction tubes which contained a 1,000-fold molar excess of unlabelled EGF (3.3×10^{-6} M) and did not exceed 4% of the total binding. The amount of ^{125}I-EGF bound to the membranes was determined by rapidly filtering the reaction mixture on Millipore EGWP filters, washing each filter with a total of 8 ml of 20 mM HEPES, *p*H 7.4, containing 0.1% BSA. The amount of radioactivity retained on each filter was determined by counting in a Nuclear Chicago γ-spectrometer. Storage of A-431 membranes at −70 °C for 2–3 months did not appreciably decrease the binding activity.

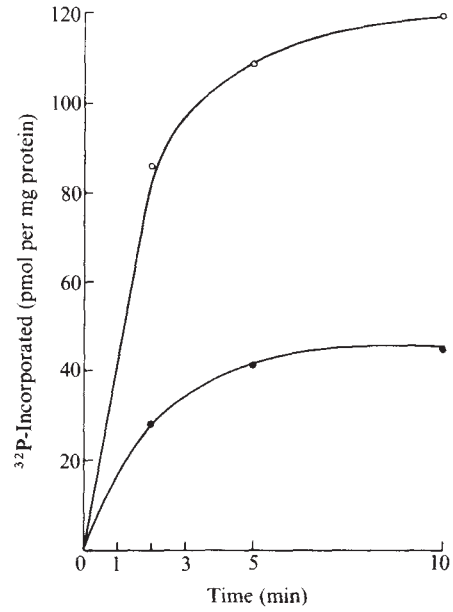


Fig. 2 Stimulation by EGF of the incorporation of ^{32}P-phosphate from [γ-^{32}P]ATP into cell membranes. The reaction mixtures contained A-431 membrane (27 μg protein), HEPES buffer (20 mM, *p*H 7.4), $MnCl_2$ (1 mM), [γ-^{32}P]ATP (15 μM, 8×10^5 c.p.m.), EGF (40 ng) and BSA (7.5 μg) in a final volume of 60 μl. The reaction tubes were placed on ice and preincubated for 10 min in the presence (○) or absence (●) of EGF. The reaction was initiated by the addition of labelled ATP and incubation at 0 °C was continued for the indicated times. The reaction was terminated by pipetting 50-μl aliquots on to squares (2 cm) of Whatman 3 MM filter paper and immediately dropping the paper into a beaker of cold 10% TCA containing 0.01 M sodium pyrophosphate. The filter papers were washed extensively with pyrophosphate-containing 10% TCA at room temperature, extracted with alcohol and ether, dried, and the radioactivity measured in a Nuclear Chicago gas-flow counter.

The specific binding of ^{125}I-labelled EGF at 0 °C to membranes prepared from A-431 tumour cells is shown in Fig. 1. In these conditions, the initial rate of binding is rapid (8 pmol of ^{125}I-EGF are bound per min per mg protein) and equilibrium is reached within 10 min. At equilibrium, approximately 18% of the hormone is membrane-bound at a specific binding activity of 9.0 pmol of EGF per mg protein. A-431 membranes, therefore, contain a concentration of EGF receptors comparable to the high levels detected in membranes prepared from human placenta[2].

Aliquots of the A-431 membrane preparation were used to determine whether the binding of EGF to membrane receptors was followed by a biochemical alteration of the membranes. As shown in Fig. 2, the addition of EGF to A-431 membranes produced a three- to fourfold stimulation of the phosphorylation of endogenous membrane proteins at 0 °C as judged by the transfer of ^{32}P-phosphate from γ-labelled ATP to membrane material. Figure 2 also shows that in the presence of EGF the initial rate of phosphorylation is increased from 16.7 to 50.0 pmol of ^{32}P-phosphate incorporated per min per mg protein. Membrane protein phosphorylation reached a maximal level by 10 min at 0 °C and a threefold higher level of ^{32}P-phosphate incorporation was achieved in the presence of EGF at all times examined.

We have partially characterised this EGF-stimulated reaction in isolated A-431 membranes and here are summarising our results, the details of which will be presented elsewhere. The ^{32}P incorporated into the membrane is trichloroacetic acid (TCA)-insoluble and only a small amount (20%) is removed by extractions with alcohol, ether and chloroform–methanol. Partial acid

© Macmillan Journals Ltd 1978

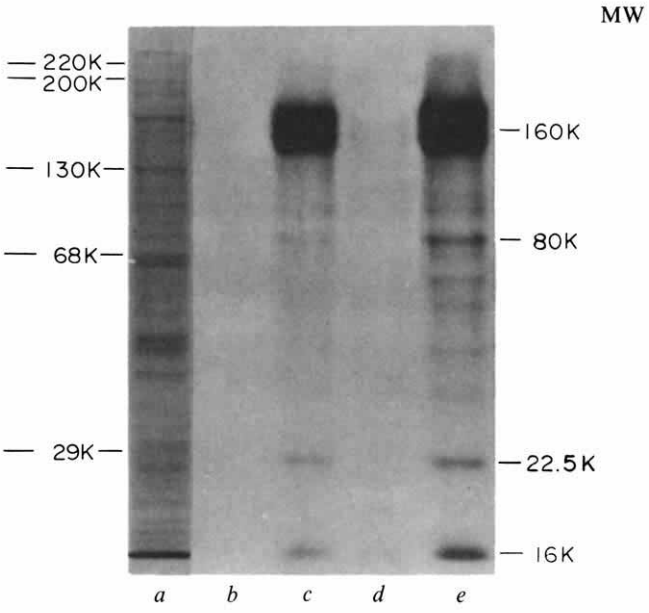


Fig. 3 Electrophoresis and autoradiography of A-431 membrane components incubated with [γ-^{32}P]ATP in the presence and absence of EGF. The phosphorylation of A-431 membranes was carried out at 0 °C in 40-μl reaction mixtures containing 30 μg of A-431 membranes, 20 mM HEPES, *p*H 7.4, 2 mM $MnCl_2$, 0.01% BSA and 15 μM [γ-^{32}P]ATP (20 Ci $mmol^{-1}$). EGF was added at a final concentration of 14×10^{-8} M immediately before the addition of ^{32}P-ATP, which initiated the reaction. The reaction was stopped after 1 or 5 min at 0 °C by the addition of 40 μl of Laemmli SDS sample buffer[8]. The mixture was boiled for 5 min and then subjected to SDS–gel electrophoresis (7.5% acrylamide) according to the method of Laemmli[8]. The gels were stained with Coomassie blue, vacuum dried and autoradiography carried out by exposing the gel to Kodak RP Royal X-O mat film for 2 d. Molecular weight (MW) standards used for calibration were erythrocyte spectrin, 240,000 and 220,000; skeletal muscle myosin, 200,000; β-galactosidase, 130,000; BSA, 68,000; carbonic anhydrase, 29,000. *a*, Coomassie blue stain of A-431 membrane components (30 μg of protein applied to the gel); *b–e*, autoradiography of A-431 membrane components phosphorylated in absence (*b*, *d*) and presence (*c*, *e*) of EGF for 1 min (*b*, *c*) or 5 min (*d*, *e*).

hydrolysis of the phosphorylated membranes and electrophoresis of the resultant mixture show that in both the control and EGF-dependent reaction the major phosphorylated product is phosphothreonine, indicating that the observed phosphorylations are directed mainly towards protein substrates. Both the endogenous membrane phosphorylation and the EGF-stimulated reaction depend on the presence of Mg^{2+} or Mn^{2+}; Ca^{2+} is ineffective. Maximal stimulation of membrane phosphorylation occurs in the presence of 3×10^{-8} M EGF. The presence of cyclic AMP or cyclic GMP (10^{-4}–10^{-6} M) has no discernible effect on the reaction. Both the basal and EGF-stimulated reaction are completely destroyed by heating the membrane for 10 min at 50 °C before the addition of [γ-^{32}P]ATP and EGF. As expected, the addition of unlabelled ATP to the reaction mixture diminishes the extent of phosphorylation. The specificity of the EGF-dependent reaction was examined by the addition of a wide variety of peptide hormones to the reaction mixture in place of EGF. We did not detect any effect of insulin, thyrotropin, follicle-stimulating hormone, growth hormone, glucagon, prolactin or luteinising hormone on the overall phosphorylation reaction. The addition of EGF to the A-431 membrane preparation seems to stimulate a protein kinase activity capable of using exogenous substrates as well as the endogenous membrane substrates. Several proteins, including histones and ribonuclease are phosphorylated when they are added to this system and their phosphorylation is stimulated two- to fourfold by the addition of EGF.

The membrane components which were phosphorylated *in vitro* in the presence and absence of EGF were analysed by SDS–polyacrylamide gel electrophoresis and autoradiography. The results shown in Fig. 3 indicate that during the first minute of incubation in the presence of EGF, membrane components having molecular weights of ~170,000 and ~150,000 are heavily phosphorylated. With continued incubation, additional membrane components are phosphorylated to a greater extent in the presence of EGF, most noticeably components with apparent MW of 80,000 and 22,500. The results also show that membrane components for which phosphorylation is increased in the presence of EGF seem to be phosphorylated in the absence of the growth factor, but to a lesser extent. Of course, these data do not show whether different sites on the same membrane proteins are phosphorylated in the presence and absence of EGF.

Thus, we have been able to demonstrate for the first time the modulation of a biochemical reaction in a cell-free membrane preparation as a consequence of the formation of EGF–receptor complexes. Following the binding of EGF, the capacity of these membranes to phosphorylate endogenous and exogenous proteins is markedly enhanced. The mechanism by which EGF enhances phosphorylation is not known; the receptor may possibly be a protein kinase, a regulatory subunit of a protein kinase or a regulator of phosphoprotein phosphatase. However, the observed net increase in phosphorylation may be an even more indirect consequence of the formation of EGF–receptor complexes. As neither the identity nor function of the phosphorylated membrane components is known, it is not possible to relate these findings to specific metabolic alterations, such as increased transport of nutrients and ions or induction of DNA synthesis, that are induced in intact cells by EGF. We have previously shown that after the binding of EGF to intact cells, this mitogen, probably as a hormone–receptor complex, is internalised by cytoplasmic vesicles[3,6]. It is not known, however, at what stage or stages in the binding and internalisation of EGF intracellular signals are generated which regulate cell growth. We suggest that the phosphorylation of membrane or membrane-associated components may be an initial event in the generation of these important signals. The effect of external stimuli or protein phosphorylation has been noted in many systems *in vivo*[9], and with insulin, by the use of sarcolemma-enriched membrane preparations *in vitro*[10].

We acknowledge the support of Biomedical Research Support Grant RR-05424-16 (G.C. and L.K.), Clinical Investigator Award from the Veteran's Administration (L.K.) and USPHS grant HD-00700 (S.C.). S.C. is an American Cancer Society Research Professor. We thank Katherine Chen, Linda Coats, Marty Reich and Suzanne Rodell for technical assistance.

GRAHAM CARPENTER*†
LLOYD KING, JR†
STANLEY COHEN*

* *Department of Biochemistry,*
and
† *Division of Dermatology, Department of Medicine,*
Vanderbilt University School of Medicine
and
Veterans Administration Hospital,
Nashville, Tennessee 37232

Received 10 July; accepted 4 September 1978.

1. Carpenter, G. & Cohen, S. in *Biochemical Actions of Hormones*, Vol. 5 (ed. Litwack, G.) (in the press).
2. O'Keefe, E., Hollenberg, M. D. & Cuatrecasas, P. *Archs Biochem. Biophys.* **164**, 518–526 (1974).
3. Haigler, H., Ash, J. F., Singer, S. J. & Cohen, S. *Proc. natn. Acad. Sci. U.S.A.* (in the press).
4. Fabricant, R. N., DeLarco, J. E. & Todaro, G. J. *Proc. natn. Acad. Sci. U.S.A.* **74**, 565–569 (1977).
5. Thom, D., Powell, A. J., Lloyd, C. W. & Rees, D. A. *Biochem. J.* **168**, 187–194 (1977).
6. Carpenter, G. & Cohen, S. *J. Cell Biol.* **71**, 159–171 (1976).
7. Bradford, M. M. *Analyt. Biochem.* **72**, 248–254 (1976).
8. Laemmli, V. K. *Nature* **227**, 680–685 (1970).
9. Greengand, P. *Science* **199**, 146–152 (1978).
10. Walaas, O. *et al. FEBS Lett.* **80**, 417–422 (1977).

Tubulin microheterogeneity increases with rat brain maturation

MICROTUBULES are present in all eukaryotic cells and have been found to have a variety of structural and dynamic roles in cell shape, division, motility, transport and secretion[1]. In nervous tissue neurite outgrowth and axoplasmic transport are also thought to depend on microtubule integrity[2]. The microtubule subunit protein, tubulin, is a heterodimer composed of two polypeptides α and β (refs 3, 4). The α and β subunits show microheterogeneity and both have been resolved into two or three components[5–8]. The question therefore arises as to whether changes occur in the relative proportions of the multiple forms of tubulin upon assumption of different roles within the nerve cell. We show here that cytoplasmic rat brain tubulin, as resolved by isoelectric focusing, is highly heterogeneous. Moreover, tubulin microheterogeneity seems to be developmentally determined, increasing from seven to nine distinct components during early postnatal rat brain maturation.

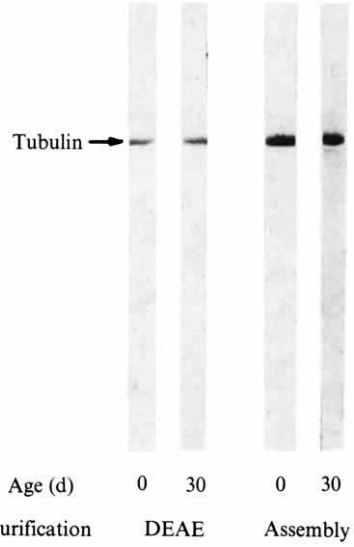


Fig. 1 SDS–polyacrylamide gel electropherogram of developing rat brain tubulin. Rat brain tubulin was prepared using DEAE–cellulose column chromatography following the procedure of Eipper[15] with slight modifications. Freshly obtained brains (2 g) from either newborn or 1-month-old rats were homogenised using a Dounce homogeniser, in 4 ml of 0.24 M sucrose, 2.5 mM $MgCl_2$ and 50 mM NaPPi, *p*H 7.0. The homogenate was centrifuged for 30 min at 30,000*g* and the supernatant obtained was loaded on to a DEAE–cellulose column (0.5 × 4 cm). After extensive washing with 0.1 M NaCl, 50 mM NaPPi, *p*H 7.0, and 2.5 mM $MgCl_2$, the tubulin peak was collected by elution with 0.26 M NaCl, 50 mM NaPPi, *p*H 7.0, and 2.5 mM $MgCl_2$, then dialysed against excess of H_2O and lyophylised. *In vitro* assembled tubulin was obtained by two cycles of polymerisation in the presence of 4 M glycerol as described by Shelanski *et al.*[16]. The purified tubulin preparations were dissolved in SDS sample buffer to yield a final concentration of 0.2 mg ml^{-1} protein, 10% glycerol, 5% 2-mercaptoethanol, 3% SDS, 0.0625 M Tris-HCl, *p*H 6.8, and 0.001% bromphenol blue and heated at 100 °C for 5 min. 10-μl aliquots were electrophoresed at 3 V cm^{-1} at 20 °C for 14–16 h on 0.75 mm thick polyacrylamide slab gels, containing 10–20% polyacrylamide gradient and 0.1% SDS[17,18]. The gels were stained with Coomassie brilliant blue.

However, extensive tubulin microheterogeneity is prominent in the brain, as tubulin from other organs is less heterogeneous.

Cytoplasmic tubulin from rat brain was identified by gel electrophoresis, its ability to undergo self-assembly, its purification by DEAE–cellulose column chromatography and

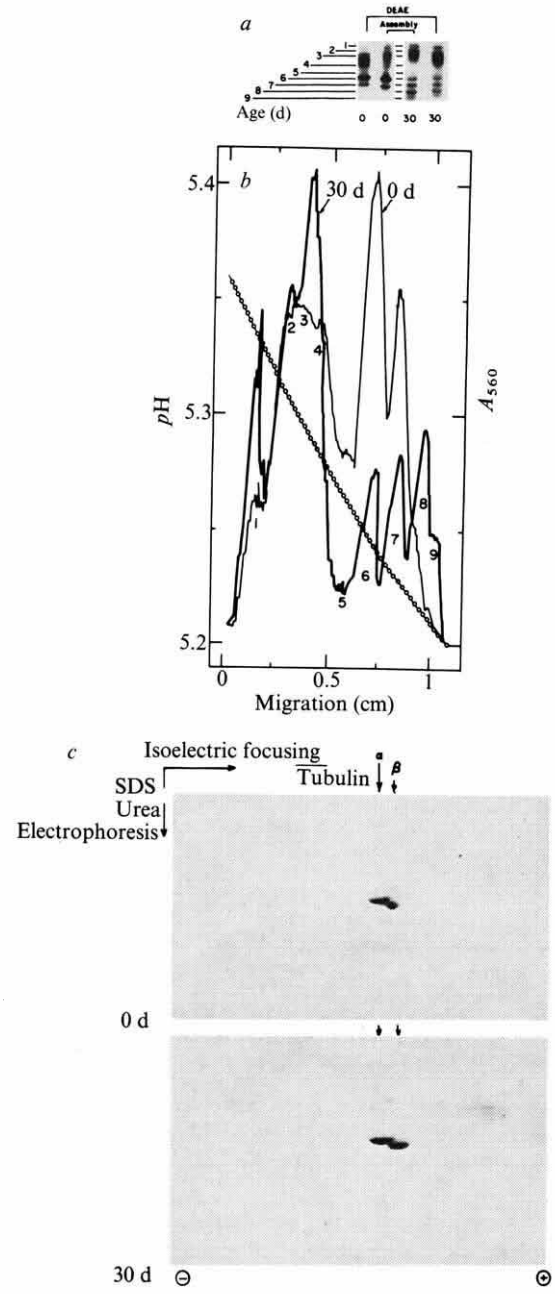


Fig. 2 Isoelectric focusing electropherogram of tubulin from newborn and 1-month-old rat brains. *a*, The tubulin preparations were obtained as described in the legend to Fig. 1. For isoelectric focusing the samples were dissolved in 9.5 M urea, 2%(v/v)NP-40, 2% ampholines (LKB) (comprised of 1.6% *p*H range 5–8 and 0.4% *p*H range 3–10) and 5% 2-mercaptoethanol and a final protein concentration of about 1 mg ml^{-1}. 10 μl aliquots were applied on to isoelectric focusing gels containing 1.6% ampholines in the *p*H range of 5–8 and 0.4% ampholines in the *p*H range of 3–10. Electrophoresis was carried out as described elsewhere[9]. On termination of the run, the gels were fixed for 60 min with 50% ethanol–7% acetic acid, soaked for 60 min with 20% methanol–7% acetic acid, stained for 5 min with 0.2% Coomassie brilliant blue–50% methanol–7% acetic acid and then destained with 20% methanol–7% acetic acid. *b*, Densitometric traces of the photographic transparencies of the isoelectric focusing gels of the DEAE–cellulose-purified tubulin were obtained by scanning at 560 nm in a Gilford 2400S spectrophotometer with a Gilford scanning attachment. The *p*H gradient was determined on termination of the electrophoresis (0). *c*, The DEAE–cellulose-purified tubulin separated by isoelectric focusing was subjected to 0.1% SDS–8 M urea–8.75% PAGE on a second dimension[9].

© Macmillan Journals Ltd 1978

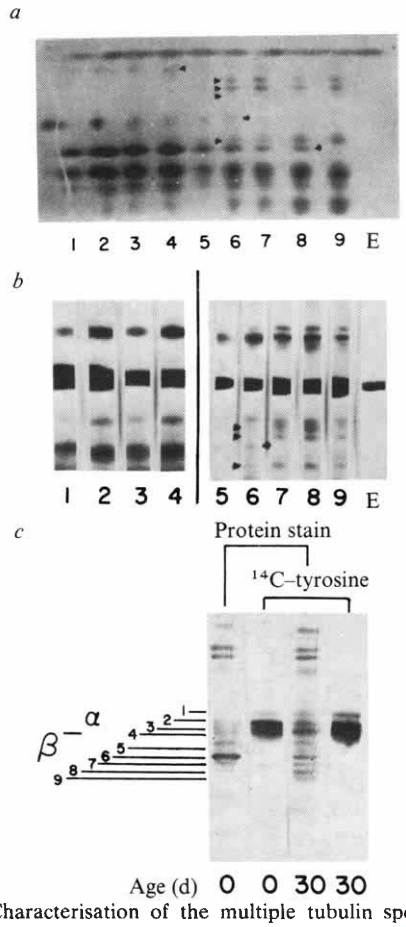


Fig. 3 Characterisation of the multiple tubulin species. *a*, *S. aureus* protease cleavage peptides of DEAE–cellulose-purified tubulin analysed on 15% polyacrylamide gel. The multiple tubulin species from 30-d-old rat brain were characterised by digestion with *S. aureus* v8 protease (36-900-1, Miles). The peptide analysis was carried out on a 15% polyacrylamide slab gel containing 0.1% SDS[19]. Tubulin bands were excised from the stained isoelectric focusing gels and soaked for 30 min in buffer containing 0.125 M Tris-HCl, *p*H 6.8, 0.1% SDS and 1 mM EDTA. The gel slices were inserted into sample slots of 15% polyacrylamide slabs with a 4-cm high upper gel. The slots were then filled with 10 μl of this buffer containing 20% glycerol and 0.001% bromphenol blue. Finally, 10 μl of the same buffer was added containing 10% glycerol and 750 ng of *S. aureus* protease. Constant voltage of 100 V was applied so that the tracking dye took 1.5 h to reach the bottom of the upper gel. The electrophoresis was then stopped for 60 min and resumed at 125 V at 10 °C and terminated 60 min after the tracking dye left the bottom of the gel. The arrow heads indicate differences in specific peptides. The isotubulins are marked in numbers according to their migration on isoelectric focusing gels. E = enzyme alone. *b*, Chymotrypsin cleavage peptides of DEAE–cellulose-purified tubulin as analysed on 15% polyacrylamide gel. The multiple tubulin species from 30-d-old rat brains were characterised by digestion with chymotrypsin (CDI-1450, Worthington). The peptide mapping was done as described above with the following modifications: 20 μg chymotrypsin were applied in each slot and the electrophoresis was stopped for 45 min instead of 60 min. The arrow heads indicate differences in specific peptides. The upper band was not marked, as its quantity varied in different experiments. The isotubulins are marked in numbers according to their migration on isoelectric focusing gels. E = enzyme alone. *c*, Tyrosylation of tubulin. Newborn and 30-d-old rat brain extracts were incubated for 10 min in the presence of L-[U^{14}C]tyrosine (50 μCi μmol^{-1}, Radiochemical Centre), as described elsewhere[11]. The samples were diluted in 9.5 M urea, 2%(v/v) NP-40 and 2% ampholine comprised of 1.6% *p*H range 5–8 and 0.4% *p*H range 3–10 and 5% 2-mercaptoethanol. 50 μl aliquots of the suspension (final concentration ~1 mg ml^{-1}) were analysed by isoelectric focusing as described in Fig. 2, followed by autoradiography of the dried gels using X-ray film (Kodak RP-54). Note that the crude homogenate from newborn rat brain contained small amounts of band 8 which was not apparent when purified tubulin was subjected to isoelectric focusing.

by vinblastine precipitation. Purified rat brain tubulin showed only one major band of about 55,000 molecular weight on electrophoresis on SDS–polyacrylamide gels (Fig. 1). In contrast, several polypeptide chains could be resolved by isoelectric focusing. Figure 2*a*, *b* shows that tubulin obtained from newborn rat brain cytoplasm was resolved into seven bands, whereas that obtained from 30-d-old rat brain displayed nine bands. It should be noted that in both cases the resolution of bands 2, 3 and 4 was less pronounced than that of the other bands. Depending on the method of purification the relative proportion of the bands varied. Thus, bands 5 and 9 were more prominent with tubulin isolated by self-assembly, than with tubulin purified by DEAE–cellulose chromatography. However, the age-dependent increase in tubulin microheterogeneity was independent of the method of purification used. The absence of bands 8 and 9 from the young rat brain tubulin did not result from an instability factor manifesting itself during homogenisation. This was shown by purifying tubulin by DEAE–cellulose chromatography, from a 1 : 1 mixture of brain from newborn and 1-month-old rats. The purified material showed bands 8 and 9 in the proportions expected from such a mixture. Moreover, isolation of tubulin in the presence of protease inhibitors such as 0.4 mM phenylmethylsulphonyl fluoride, or a mixture of 40 μg ml^{-1} leupeptin, 20 μg ml^{-1} pepstatin A and 1 mg ml^{-1} bacitracin, did not change the pattern, indicating that the microheterogeneity does not arise from artefactual proteolysis (see also below). When the ampholine polyacrylamide gels were subjected to 0.1% SDS–8 M urea–8.75% polyacrylamide gel electrophoresis (PAGE) on a second dimension[9], bands 1, 2, 3 and 4 were found to correspond to the α-tubulin subunit and bands 5, 6, 7, 8 and 9 to the β subunit[10] (Fig. 2*c*). Thus, isoelectric focusing did not alter the tubulin mobility on electrophoresis on urea–SDS–polyacrylamide gels.

The relationship of the various tubulin bands was examined by protease digestion. After isoelectric focusing of tubulin extracted from 30-d-old rat brain the different bands were excised, subjected to *Staphylococcus aureus* protease or chymotrypsin digestion and analysed by SDS–PAGE. The results in Fig. 3*a*, *b* show a close similarity in the overall distribution of the polypeptide chains derived from each preparation, indicating a structural homology between the isotubulins and eliminating the possibility of non-tubulin contaminants. The distribution pattern of the cleavage peptides was independent of the proteolytic enzyme concentration. Moreover, comparison of the peptide maps of tubulin with non-related proteins such as actin showed marked differences (data not shown). Closer examination of the peptide maps (Fig. 3*a*, *b*) reveals differences between bands 1–5 and 6–9. In addition, band 9 differs from bands 6–8 (Fig. 3*a*) and band 6 differs from bands 7–9 (Fig. 3*b*). Although band 5 closely resembles bands 1–4 (α-tubulin), it migrates as β-tubulin when resolved by two-dimensional electrophoresis[10]. Band 5 also differs from bands 1–4 in its inability to undergo *in vitro* tyrosylation (Fig. 3*c*); this reaction is characteristic of the α-tubulin subunit[11]. All the above results suggest the existence in adult rat brain of multiple non-identical tubulin species exhibiting different isoelectric points, but similar molecular weights.

Extensive tubulin microheterogeneity seems to be prominent in the brain, as tubulin from other tissues such as 30-d-old rat liver and spleen as well as neuroblastoma–glioma NG108-15 hybrid cells were less heterogeneous and contained two major subunits (Fig. 4*a*, *b*). Peptide map analysis showed that the two tubulin species were similar to the corresponding isotubulins of rat brain, indicating no major contamination in this preparation (data not shown). The pronounced microheterogeneity of rat brain tubulin is not due to specific proteolytic cleavage occurring during the homogenisation and isolation procedures. This was shown by purifying tubulin from a mixture of ^{35}S-methionine-labelled neuroblastoma–glioma hybrid cells in the presence of 30-d-old rat brain. Figure 4*b* shows that such a mixing experiment did not produce additional radioactive isotubulins. In

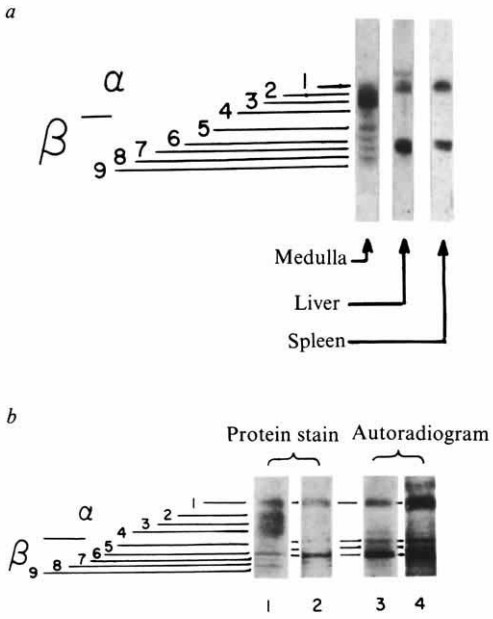


Fig. 4 Multiple forms of tubulin in various tissues. *a*, Isoelectric focusing of 30-d-old rat tubulin prepared from medulla oblongata, liver and spleen. Tubulin was purified by precipitation with vinblastine (Lilly). The various tissues were homogenised in 2 volumes of 0.25 M sucrose, 10 mM phosphate buffer, *p*H 7.4, 10 mM $MgCl_2$ and 0.2 mM GTP, and the supernatants (1 mg protein ml^{-1}) obtained after 1 h centrifugation at 105,000*g* were subjected to vinblastine (1 mg ml^{-1}) precipitation[18]. The precipitated proteins were dissolved in 9.5 M urea, 2%(v/v)NP-40, 2% ampholines (comprising 1.6% *p*H range 5–8 and 0.4% *p*H range 3–10) and 5% 2-mercaptoethanol, and a final protein concentration of ~1 mg ml^{-1}. 10 μl aliquots were applied on to isoelectric focusing gels containing 1.6% ampholines in the *p*H range of 5–8 and 0.4% ampholines in the *p*H range of 3–10. The gels were stained with Coomassie brilliant blue (as described in the legend to Fig. 2). *b*, Isoelectric focusing of tubulin from neuroblastoma–glioma hybrid cells prepared in the presence of 30-d-old rat brains. Neuroblastoma–glioma hybrid cells (NG108-15) grown as described elsewhere[20] were incubated for 24 h in the presence of ^{35}S-methionine (700 Ci $mmol^{-1}$, Radiochemical Centre), 100 μCi per 10-cm diameter plates containing 10 ml medium. The packed cells (100 μl of a pellet centrifuged at 1,000*g* for 5 min) were homogenised in 200 μl of 0.25 M sucrose, 10 mM phosphate buffer, *p*H 7.4, 10 mM $MgCl_2$ and 0.2 mM GTP, in one experiment. In a second experiment 100 μl of packed labelled cells were mixed with 0.1 g of 30-d-old rat brain and homogenised in 400 μl of the above buffer. The 105,000*g* supernatants obtained after 60 min centrifugation were subjected to vinblastine precipitation[18]. The vinblastine pellets were analysed by isoelectric focusing followed by autoradiography (as described in the legends to Figs 2, 3). 1, Tubulin prepared from a mixture of hybrid cells and brains stained with Coomassie brilliant blue; 2, tubulin prepared from hybrid cells stained with Coomassie brilliant blue; 3, autoradiogram of tubulin from hybrid cells prepared in the presence of rat brain homogenate; 4, autoradiogram of tubulin isolated from hybrid cells.

addition, the limited tubulin microheterogeneity found in neuroblastoma–glioma hybrid cells is not the result of a specific protease inhibitor. As shown in Fig. 4*b*, there was no reduction in the multiple species of brain tubulin isolated in the presence of neuroblastoma–glioma hybrid cells (Coomassie brilliant blue-stained bands).

In conclusion, an increase in the tubulin microheterogeneity was demonstrated during brain maturation. This process coincides with an increase in the ability of tubulin to undergo self-assembly[12–14], suggesting that the changes in the multiple tubulin forms control microtubule formation.

We thank Mr M. D. Walker and Dr J. Prives for discussions, Mrs C. Kalcheim for help in the tyrosylation experiment and Mrs A. Zutra for the neuroblastoma–glioma cells. This work was supported in part by grants (to U.Z.L) from the US-Israel Binational Science Foundation (BSF), Jerusalem.

ILLANA GOZES
URIEL Z. LITTAUER

Department of Neurobiology, Weizmann Institute of Science, Rehovot, Israel

Received 18 August; accepted 9 October 1978.

1. Olmsted, J. B. & Borisy, G. G. *A. Rev. Biochem.* **42**, 507–540 (1973).
2. Shelanski, M. L. & Feit, H. in *The Structure and Function of the Nervous Tissue*, Vol. 6 (ed. Bourne, G. H.) 47–80 (Academic, New York, 1972).
3. Bryan, J. & Wilson, L. *Proc. natn. Acad. Sci. U.S.A.* **68**, 1762–1766 (1971).
4. Luduena, R., Wilson, L. & Shooter, E. M. *J. Cell Biol.* **63**, 202a (1974).
5. Feit, H., Slusarek, L. & Shelanski, M. L. *Proc. natn. Acad. Sci. U.S.A.* **68**, 2028–2031 (1971).
6. Witman, G. B., Carlson, K. & Rosenbaum, J. L. *J. Cell. Biol.* **54**, 540–555 (1972).
7. Feit, H. & Shelanski, M. L. *Biochem. biophys. Res. Commun.* **66**, 920–927 (1975).
8. Bibring, T., Baxandall, J., Denslow, S. & Walker, B. *J. Cell Biol.* **69**, 301–312 (1976).
9. O'Farrell, P. H. *J. biol. Chem.* **250**, 4007–4021 (1975).
10. Gozes, I. & Littauer, U. Z. in *Proc. 6th Int. Soc. Neurochemistry* (eds Sourkes, T. L. & Clausen, J.) 121 (Copenhagen, Denmark, 1977).
11. Raybin, D. & Flavin, M. *Biochem. biophys. Res. Commun.* **65**, 1088–1095 (1975).
12. Fellous, A., Francon, J., Virion, A. & Nunez, J. *FEBS Lett.* **57**, 5–8 (1975).
13. Fellous, A., Francon, J., Lennon, A. M. & Nunez, J. *FEBS Lett.* **64**, 400–403 (1976).
14. Schmitt, H., Josephs, R. & Reisler, E. *Nature* **265**, 653–655 (1977).
15. Eipper, B. A. *Proc. natn. Acad. Sci. U.S.A. 69*, 2283–2287 (1972).
16. Shelanski, M. L., Gaskin, F. & Cantor, C. R. *Proc. natn. Acad. Sci. U.S.A.* **70**, 765–768 (1973).
17. Maizel, J. V., Jr in *Methods in Virology*, Vol. 5 (eds Maramorosch, K. & Koprowski, H.) 179–246 (Academic, New York, 1971).
18. Gozes, I., Schmitt, H. & Littauer, U. Z. *Proc. nam. Acad. Sci. U.S.A.* **72**, 701–705 (1975).
19. Cleveland, D. W., Fischer, S. G., Kirschner, M. W. & Laemmli, U. K. *J. biol. Chem.* **252**, 1102–1106 (1977).
20. Klee, W. A. & Nirenberg, M. *Proc. natn. Acad. Sci. U.S.A.* **71**, 3474–3478 (1978).

Fibroblasts degrade newly synthesised collagen within the cell before secretion

COLLAGEN, the major extracellular protein of most tissues, is considered a relatively inert macromolecule, with a turnover rate slower than that of most proteins[1]. There is, however, a portion of the total collagen pool that is rapidly turning over, for tracer studies of collagen metabolism in intact animals have demonstrated that a significant amount of newly synthesised collagen is apparently degraded within hours of its synthesis[2]. Further, direct evidence for the rapid degradation of collagen has been provided by studies of rabbit lung explants which showed that 30–40% of newly synthesised collagen was degraded within minutes of its synthesis[3]. This phenomenon is of interest because it involves such a large proportion of collagen production, and thus may significantly influence the quantity of collagen within a tissue and hence organ structure and function. The time course of this rapid degradation of collagen suggests that it occurs within the collagen-producing cells before its secretion. Although it has been repeatedly demonstrated that intracellular degradative processes have a central role in regulating the quantity of proteins that function within cells, intracellular degradation has not been considered a general mechanism for regulating the quantity of proteins destined for extracellular function[4–6]. We describe here a direct evaluation of the hypothesis that extracellular levels of collagen are regulated, in part, by the rapid degradation of newly synthesised collagen before it reaches the extracellular space. We have studied the size and kinetics of collagen production by fibroblasts in culture, and have found that not only is a significant proportion of newly synthesised collagen degraded intracellularly, but also that the process can be modulated by manipulating the structure of the collagen molecule.

All studies used subcultivations 10–20 of a human fetal lung fibroblast termed HFL-1; this cell strain is diploid, free of

0028–0836/78/0276—0413$01.00

© Macmillan Journals Ltd 1978

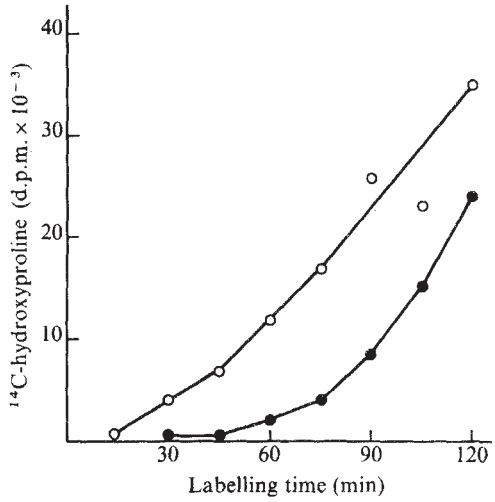


Fig. 1 Time course of the synthesis and secretion of collagen by HFL-1 fibroblasts. Collagen is represented as the quantity of ^{14}C-hydroxyproline found in the cell layer (○) or media (●) at various times following addition of ^{14}C-proline to the cultures. Cells were plated at 10^6 cells per 150-mm culture dish and maintained at 37°C (10% CO_2 in air) in 20 ml growth medium (Dulbecco's modified Eagle's medium (DMEM) with added fetal calf serum (10%), glutamine (0.06%), penicillin (100 U ml^{-1}) and streptomycin (100 μg ml^{-1})) which was changed every 3 d. All studies of collagen production were done after the cells had reached the stationary phase (usually 7–8 d after subcultivation). 24 h before labelling, ascorbic acid (50 μg ml^{-1}) was added to the growth medium. To evaluate collagen production by HFL-1 fibroblasts, the cells were washed (×3) with phosphate-buffered saline (PBS) *p*H 7.4, and incubated with growth medium supplemented with ascorbic acid (50 μg ml^{-1}) and β-aminopropionitrile (50 μg ml^{-1}). After 1 h, the supplemented growth medium was discarded and replaced with 10 ml of labelling medium (DMEM, dialysed fetal calf serum (10%), glutamine (0.06%), penicillin (100 U ml^{-1}), streptomycin (100 μg ml^{-1}), ascorbic acid (50 μg ml^{-1}), β-aminopropionitrile (50 μg ml)$^{-1}$ and ^{14}C-proline (10 μCi ml^{-1}, 260 mCi mmol^{-1}, Schwarz-Mann)). The dialysed fetal calf serum used in the labelling medium was prepared by dialysing fetal calf serum against PBS so as to remove amino acids that would dilute the tracer but to retain antiproteases. At each time point indicated, culture plates were removed from the incubator, placed on ice, and the culture medium collected. The cell layer was washed with PBS (×3) and the washings combined with the medium. The cell layers were scraped from the plate using 20 ml 0.5 M acetic acid. The media and cell layer samples were sonicated, heated at 100°C for 10 min to inactivate proteases, and aliquots taken for quantitation of ^{14}C-hydroxyproline using acid hydrolysis (6M HCl, 110°C, 20 h) and ion-exchange chromatography as previously described[3]. The values for ^{14}C-hydroxyproline represent the total disintegration per min (d.p.m.) of ^{14}C-hydroxyproline found in the cells or medium for each plate independent of whether the collagen was intact or degraded. The total number of cells per plate varied by less than 5%.

mycoplasma and viral infection, and exhibits contact inhibition[7]. HFL-1 synthesises both collagen types I and III, which together represent 3–5% of the total proteins synthesised by this cell. Degradation of newly synthesised collagen in HFL-1 was followed by labelling the cells, in the presence of fetal calf serum, with either ^{14}C-proline or ^{14}C-lysine and quantitating the proportion of newly synthesised proteins containing ^{14}C-hydroxyproline or ^{14}C-hydroxylysine that could pass through a dialysis membrane[3]. This experimental design for quantitating collagen degradation was based on the fact that in cultures of human fibroblasts, labelled hydroxyproline and hydroxylysine can be used as specific markers for newly synthesised collagen. Although both amino acids are found in Clq, a complement component also synthesised by the fibroblast, estimates of the relative synthetic rates for Clq and collagen demonstrate that the Clq is synthesised at 0.1% the rate (per fibroblast) of collagen[8].

As expected for a protein destined for extracellular function, newly synthesised collagen was detected in the cell layer of the HFL-1 cultures 15–30 min before its appearance in the culture medium (Fig. 1). The time course for collagen synthesis and secretion found in these cells is comparable to that of other vertebrate mesenchymal cells[9].

Strikingly, 30 min after the addition of the tracer, even though the newly synthesised collagen had not then been secreted (Fig. 1), approximately 30% of the labelled collagen in the cell layer was already degraded to peptides small enough to pass through a dialysis membrane (Fig. 2). One hour after the addition of the tracer, degraded collagen was detected in both the cell layer and the medium, but the combined total of degraded collagen represented a constant fraction of the total amount of collagen synthesised during the labelling period. Even 24 h after addition of the tracer to the cultures, the same fraction of newly synthesised collagen was found to be degraded, although by this

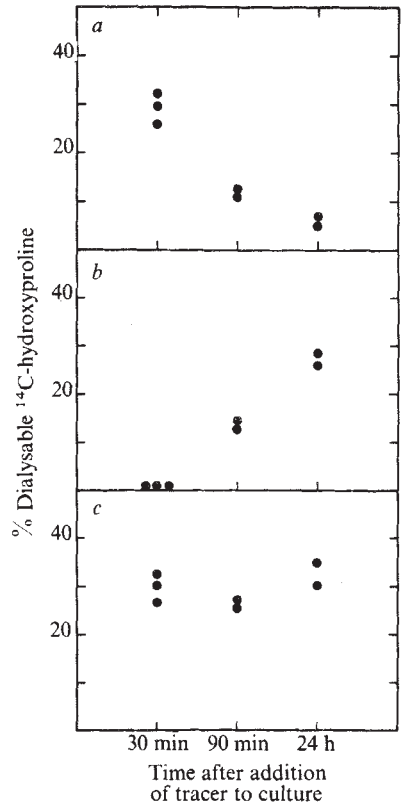


Fig. 2 Proportion of newly synthesised collagen degraded by HFL-1 cells. Cells were maintained in culture and labelled with ^{14}C-proline as described in Fig. 1. After a 30-min labelling period, the labelling medium was discarded, the cell layers were washed three times with PBS, and 20 ml of 'chase' medium (identical to labelling medium except that 1 mM unlabelled proline was substituted for ^{14}C-proline) was added. At 0, 1 or 23.5 h after addition of the chase medium (representing respectively, 30 min, 90 min or 24 h after the tracer had been originally added to the culture), the incubation was stopped and the media and cell layer collected separately as described in Fig. 1. Following sonication and heating, an aliquot was taken for analysis of ^{14}C-hydroxyproline in the *a*, cell layer, *b*, media or *c*, cell layer and media that was dialysable[3]. Per cent dialysable ^{14}C-hydroxyproline was calculated from the formula: % dialysable ^{14}C-hydroxyproline = [(d.p.m. of total ^{14}C-hydroxyproline in the sample) − (d.p.m. of ^{14}C-hydroxyproline in the sample after dialysis)] × 100/[d.p.m. of total ^{14}C-hydroxyproline in the sample]. 30 min after addition of ^{14}C-proline to the culture (that is, 9 h chase time), an average of 30% of ^{14}C-hydroxyproline in the culture was dialysable, and all was present in the cell layer. With time after the chase, the proportion of dialysable ^{14}C-hydroxyproline in the cell layer decreased and the proportion of dialysable ^{14}C-hydroxyproline in the media increased. Even so, the proportion of the total ^{14}C-hydroxyproline in the culture (cell layer and media) that was dialysable remained constant (*c*, $P > 0.1$, Student's two-tailed *t*-test).

time most of the degraded peptides had moved from the cell layer to the culture medium. Evaluation of the size distribution of the peptides derived from degraded collagen demonstrated that more than 90% of the degraded collagen was reduced to less than 500 daltons within 6 h (data not shown). Thus, not only do these cells rapidly metabolise a significant proportion of newly synthesised collagen, but they also degrade these macromolecules to very small fragments. However, a proportion of the degraded collagen was still in peptide linkage, for chromatography of the dialysable material on the amino acid analyser before acid hydrolysis resulted in a significant reduction in the ^{14}C-hydroxyproline recovered (R. Berg and R. G. C., unpublished observations). This, and the fact that the proportion of dialysable ^{14}C-hydroxyproline did not increase with time of culture (Fig. 2), are consistent with the overwhelming evidence that any hydroxyproline found in these cultures is derived from proline residues in peptide linkage (rather than an ester link such as that found in prolyl-tRNA)[10–13].

When HFL-1 cells were labelled with ^{14}C-lysine in the same conditions as with ^{14}C-proline, approximately 22% of the labelled ^{14}C-hydroxylysine was found in peptides small enough to pass through a dialysis membrane (Fig. 3). This observation is important because hydroxyproline (but not hydroxylysine) residues are present in the N-terminal ('Pn') precursor piece of collagen pro α1(I) chains (D. Hörlen and P. Fietzek, personal communication), and the finding of ^{14}C-hydroxylysine in dialysable peptides indicates that the material being degraded in these cultures includes the α-chain portion of newly synthesised collagen[3].

The fact that the degraded collagen was detected in the cell layer before the time required for HFL-1 cells to synthesise and secrete collagen argues strongly that this degradative process was occurring within, and not outside, the cells. Other evidence supports this conclusion. First, at the time degraded collagen was detected in the cell layer (Fig. 2*a*) none was found in the media (Fig. 2*b*). Although it is conceivable that collagen was degraded outside the cells and the resulting fragments then internalised, the time course of collagen degradation argues against this. Second, when purified ^{14}C-labelled human type I procollagen was incubated with HFL-1 cultures for 6 h, 100% of the labelled procollagen could be recovered intact, suggesting that the culture media did not contain active proteases capable of degrading collagen (data not shown). However, to ensure that any putative extracellular collagen-specific proteases would be inhibited, all experiments were carried out in the presence of 10% fetal calf serum, a concentration sufficient for the complete inhibition of mammalian collagenase capable of digesting 12 μg of collagen per h at 37°C (ref. 14). In addition, even in the absence of the fetal calf serum, no active collagenase could be detected in the culture media of HFL-1 fibroblasts incubated for 24 h.

Third,when the cells were pulsed with ^{14}C-proline for 30 min and then chased with unlabelled proline, the proportion of labelled collagen degraded did not change even when the cultures were continued for 24 h. If degradation was due to proteolytic enzymes outside the cells, it is likely that a progressive increase in collagen degradation would have been noted as the culture continued. The time invariance of the proportion of total collagen degraded in the pulse–chase experiments also suggested that the dialysable collagen peptides found in these cultures represented active degradation rather than simply uncompleted collagen chains 'nascent' on polyribosomes at the end of the pulse period. If it was the latter, these should have been completed during the 'chase' and the proportion of dialysable ^{14}C-hydroxyproline would have decreased.

Thus, it is apparent that a large proportion of collagen produced by diploid human fibroblasts is degraded rapidly within the cells. As this degradative process occurs before the time necessary for secretion, and as the proportion of collagen that is degraded does not change with time, it is likely that this phenomenon occurs before secretion and not after a process of secretion and phagocytosis. If the latter were responsible, the time required would be much longer and the labelled collagen in the cultures would be progressively destroyed as the culture times were extended. It is not certain, however, that all the dialysable ^{14}C-hydroxyproline and ^{14}C-hydroxylysine found in these cultures were necessarily derived from intact collagen chains. If, for example, there was premature termination of some of the collagen being synthesised, these incomplete chains might be more susceptible to degradation than fully completed pro α chains.

If intracellular collagen degradation is an important post-translational mechanism which modulates the differentiated state of fibroblasts with respect to collagen production, these cells should have means available to vary the relative proportion of newly synthesised collagen that is degraded by this process.

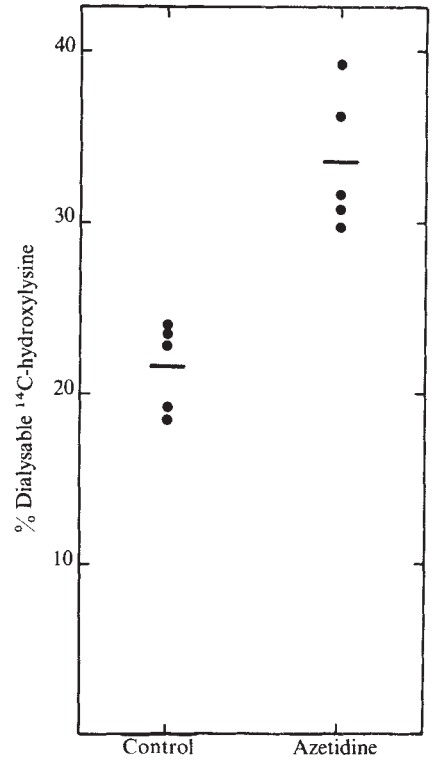


Fig. 3 Influence of azetidine-2-carboxylic acid (azetidine) on the degradation of newly synthesised collagen by HFL-1 cells. HFL-1 cells were maintained in cultur a described in Fig. 1. To evaluate collagen production in the presence of azetidine, the cells were first washed three times with PBS and incubated with lysine-free DMEM supplemented with ascorbic acid (50 μg ml^{-1}) and β-aminopropionitrile (50 μg ml^{-1}). After 1 h, the supplemented growth medium was discarded and replaced with labelling medium (lysine-free DMEM, dialysed fetal calf serum (10%), glutamine (0.06%), penicillin (100 U ml^{-1}), streptomycin (100 μg ml^{-1}), ascorbic acid (50 μg ml^{-1}), β-aminoproprionitrile (50 μg ml^{-1}) and ^{14}C-lysine (10 μCi ml^{-1}, 300 mCi mmol^{-1}, Schwarz-Mann)), to which either proline (100 μg ml^{-1}, Schwarz-Mann) or azetidine (100 μg ml^{-1}, Calbiochem) had been added. Following incubation for 12 h, the cell layers and culture media were combined, sonicated and heated at 100°C for 10 min. An aliquot was taken for the measurement of total ^{14}C-hydroxylysine in the sample and the remainder dialysed against 0.5 M acetic acid and 0.1 mM hydroxylysine. The proportion of newly synthesised collagen (represented by ^{14}C-hydroxylysine) thgt was degraded was quantitated as follows: % dialysable ^{14}C-hydroxylysine = [(d.p.m. of total ^{14}C-hydroxylysine in the sample) − (d.p.m. of ^{14}C-hydroxylysine in the sample after dialysis)] × 100/[d.p.m. of total ^{14}C-hydroxylysine in the sample]. Each experimental point represents the per cent dialysable ^{14}C-hydroxylysine observed for separate culture plates. The control plates (cells labelled with ^{14}C-lysine in the presence of proline) averaged 22 ± 3% and the azetidine plates (cells labelled with ^{14}C-lysine in the presence of azetidine instead of proline) averaged 33 ± 5%; the values were significantly different at the $P < 0.01$ level (Student's two-tailed *t*-test).

There is accumulating evidence that intracellular degradative processes can recognise and destroy protein with abnormal conformations[4], so that one way to evaluate whether cells can alter intracellular collagen degradation is to manipulate the cells to produce collagen molecules with an abnormal structure. The approach used was to incubate HFL-1 cells in medium in which azetidine-2-carboxylic acid had been substituted for proline. Azetidine, the lower homologue of proline, is readily taken up by cells, recognised by prolyl-tRNA, and incorporated into newly synthesised collagen in place of proline[15]. However, the presence of the azetidine ring in the collagen polypeptide chain precludes normal helix formation[16,17]. Interestingly, when this was done, intracellular collagen degradation increased almost twofold (Fig. 3). Thus, cells seem to have means by which they can modulate the fraction of total collagen that is degraded within the cell before secretion.

As might be expected for a macromolecule with such a critical role in organ structure and function, there seem to be several ways in which the metabolism of collagen can be regulated. For collagen already outside the cell, extracellular proteases such as collagenase are undoubtedly paramount in this regard[18]. The present study suggests that mechanisms are also available by which cells can degrade a significant, yet adjustable, proportion of newly synthesised collagen before it ever reaches the extracellular space. Further study will be required to ascertain whether this is a general control phenomenon for secreted proteins or whether it is unique to collagen.

ROBERT S. BIENKOWSKI
BRUCE J. BAUM
RONALD G. CRYSTAL

Pulmonary Branch,
National Heart, Lung and Blood Institute,
Bethesda, Maryland

Received 5 July; accepted 21 September 1978.

1. Jackson, D. S. in *Biology of Fibroblast* (eds Kulonen, E. & Pikkarainen, J.) 417–420 (Academic, New York, 1973).
2. Kivirikko, K. I. *Int. Rev. Connective Tissue Res.* **5**, 93 (1970).
3. Bienkowski, R. S., Cowan, M., McDonald, J. A. & Crystal, R. G. *J. biol. Chem.* **253**, 4356 (1978).
4. Goldberg, A. L. & Dice, J. R. *A. Rev. Biochem.* **43**, 835 (1974).
5. Goldberg, A. L. & St. John, A. C. *A. Rev. Biochem.* **45**, 747 (1976).
6. Schimke, R. T. & Katunuma, N. (eds) *Intracellular Protein Turnover* (Academic, New York, 1975).
7. Baum, B. J., McDonald, J. A. & Crystal, R. G. *Biochem. biophys. Res. Commun.* **79**, 8 (1977).
8. Reid, K. B. M. & Solomon, *Eur. Biochem. J.* **167**, 647 (1977).
9. Kao, W. W-Y., Berg, R. A. & Prockop, D. J. *J. biol. Chem.* **252**, 8391 (1977).
10. Udenfriend, S. *Science* **152**, 1335 (1966).
11. Bornstein, P. *A. Rev. Biochem.* **43**, 567 (1974).
12. Cardinale, G. J. & Udenfriend, S. *Adv. Enzym.* **41**, 245 (1974).
13. Kivirikko, K. I. & Prockop, D. J. *Archs Biochem. Biophys.* **118**, 611 (1967).
14. Horowitz, A. L. & Crystal, R. G. *Biochem. biophys. Res. Commun.* **69**, 296 (1976).
15. Lane, J. M., Dehm, P. & Prockop, D. J. *Biochem. biophys. Acta* **236**, 517 (1971).
16. Berman, H. M., McGanday, R. L., Burgner, J. W. & Van Ettern, R. L. *J. Am. chem. Soc.* **91**, 6177 (1969).
17. Baum, B. J., Johnson, L.S., Franzblau, C. & Troxler, R. F. *J. biol. Chem.* **250**, 1464 (1975).
18. Gross, J. in *Biochemistry of Collagen* (eds Ramachandran, G. N. & Reddi, A. H.) 275–317 (Plenum, New York, 1976).

Nitrogen fixation in the rhizosphere of rice

THE discovery of nitrogen-fixing activity associated with the rhizosphere of non-leguminous plants such as tropical grasses, rice and maize has provoked considerable interest[1–9]. Discrepancies in the measurements of this nitrogen-fixing activity by the acetylene reduction assay in different conditions prompted us to develop a modified assay system. We report here the use of this assay to find significant differences in the rates of acetylene reduction associated with the rhizosphere of 50 different strains of rice. This suggests that the rice plant genotype influences the association with nitrogen-fixing bacteria in its rhizosphere.

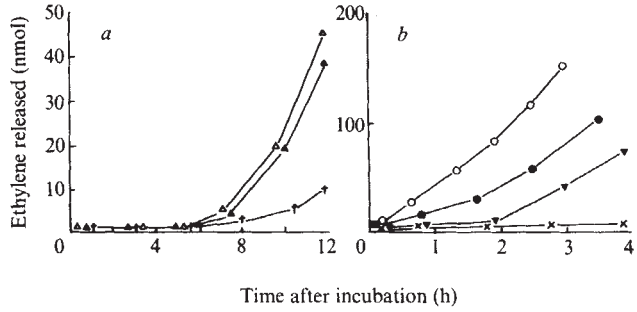


Fig. 1 Time course of C_2H_2-reducing activity in different assay conditions. C_2H_2-reducing activity was measured by gas chromatography (Shimadzu GC-3BF) with H_2-flame ionisation detector. Assay conditions were a 2-cm column containing Porapack Q of 80–100 mesh with N_2 gas as carrier at a flow-rate of 40 cm^3 min^{-1} at 80 °C. *a*, The gas phase was replaced with a mixture of 0.1 atm C_2H_2 and 0.9 atm He and incubated at 30 °C until assayed. (▲), Excised roots washed with tap water and kept in a flask, per g dry root; (△), whole plant with soil kept in a polyethylene bag, per g dry root; (+), soil alone kept in a flask, per 100 g. *b*, Whole plant kept in a vessel containing different gas phase (per g dry root). Gas phase: 0.1 C_2H_2 and 0.9 He (○); 0.1 C_2H_2 and 0.9 air (●); and after the gas phase was replaced with 1 atm air and kept for 30 min in order to expose rhizosphere to air, it was replaced again with a mixture of 0.1 C_2H_2 and 0.9 He (▼). Control, no C_2H_2 (×).

The first report of high levels of nitrogen-fixing activity associated with isolated roots of maize was by von Bülow and Döbereiner[4]. Subsequently, using the maize–*Spirillum lipoferum* system, Burris *et al.*[10] showed that the acetylene-reducing activity of intact plants or soil cores containing the bulk of a relatively undisturbed root system was far less than that of preincubated roots. Similar observations were made in rice. When the excised root system or intact plant kept in a polyethylene bag was preincubated under either aerobic or anaerobic conditions, the activity remained almost zero for the first few hours after incubation[8,11–13] and increased exponentially on further incubation. In the maize–*Spirillum* system, a marked proliferation of the bacteria during preincubation was observed[14,15]. This suggested that the activity detected after a lag period might be due to the increased number of bacteria and might not represent the activity in the intact plant.

We devised the following assay system with these observations in mind. An intact plant including the soil and water of the rice rhizosphere and the bulk of the relatively undisturbed root system was enclosed in a vessel such as a desiccator so that the exposure of the rhizosphere to air was minimised. Rigorous displacement of the gas phase in the vessel with a mixture of helium and acetylene (9 : 1 by volume) was achieved two times by an evacuation pump (up to 0.02 atm) and flushing the gas mixture up to 1 atm. The vessel was flushed with the gas mixture for 30 min to equilibrate the system, evacuated and flushed again. The system was incubated in normal laboratory light at 28 °C and fractions of the gas phase were withdrawn with a syringe for assay. The ethylene produced from acetylene was measured by gas chromatography[16,17]. It was thought that the assay procedure enabled us to displace the air remaining in the assay sample with the gas mixture containing acetylene as a substrate so that the substrate could effectively reach the inner part of the sample where nitrogen fixation was active without disturbing the rhizosphere. The observations described below suggest that the above two assay conditions must be fulfilled to measure the nitrogen-fixing activity in rhizosphere correctly.

Figure 1 shows the kinetics of acetylene reduction in the rice rhizosphere measured in the various assay conditions. When the assay system described above was used, considerable acetylene reduction activity was found with certain strains of rice. The production of ethylene from acetylene was linear and no time lag in the reaction was observed. The activity detected by this method was found in conditions in which almost no time was allowed for bacterial growth in the system. On the other hand,

© Macmillan Journals Ltd 1978

the other assay system using polyethylene bags[13] or excised roots in a flask[4,12] showed a distinct time lag before ethylene production was detected. When acetylene was not added to the system, no ethylene was produced. Soil without plants showed no detectable acetylene reducing activity during a 6-h incubation and later small but significant amounts of ethylene were observed. Aerobic conditions for acetylene reduction gave lower activity than anaerobic conditions. Exposure of the rhizosphere to the air for 30 min also reduced activity considerably. Plants were incubated in the light and it is possible that photo-produced oxygen is altering the physiological environment of the rhizosphere to favour microaerophilic bacteria. This, however, is unlikely as our whole plant assay system showed linear acetylene reduction without a lag. From this we deduce that our assay system measures functional nitrogenase activity in the rhizosphere and provides a useful method to assay the nitrogen-fixing activity of the plant–rhizosphere complex.

The nitrogen-fixing activity of the rhizosphere of 50 strains of rice was then measured. The strains used were collected from various parts of tropical Asia and preserved at the Genetic Stocks Centre, National Institute of Genetics, Mishima. The plants were grown in pots until they flowered and the activity of these intact plants, two of each strain, was then measured in our assay conditions. Measurements were taken 1 and 2 h after incubation and the amount of ethylene produced per hour per plant was calculated. The values obtained early after incubation reflected the activity of the rhizosphere as then any activity caused by the growth of nitrogen-fixing bacteria could be ruled out. The rate of nitrogen fixation varied widely between strains (Table 1). In addition, 20 pots containing soil without plants were measured as controls. No activity was detected in these pots during the first few hours of incubation at room temperature. Results shown in Table 1 were measured within 2 h of incubation. There was, therefore, a distinct difference in the nitrogen-fixing activities between controls and experimental pots. The variation among strains was statistically significant (Table 2). Activity expressed per g dry root per hour also showed a wide range of significant variation between strains. The highest value was about 1,100 nmol per plant per hour. Strains showing high activity were collected in India and Thailand. However, the value was rather low compared with those (2,000–6,000 nmol per g dry root per h) obtained by Dommergues *et al.*[9] using rice strain IR 8. Also using IR 8, Lee *et al.*[13] reported lower values (1,870 nmol per pot per h). Because the soil used in those experiments and experiments reported here might have contained different microbial floras and the activity was also measured in different assay conditions, it is difficult to compare these results.

There was a correlation ($r = 0.895$, significant at the 1% level) between root weight and nitrogen-fixing activity within a strain. This means the larger the amount of root per unit area, the higher the activity. Roots are, therefore, apparently essential for bacterial nitrogen fixation in the paddy field.

Seeds of a rice strain associated with high nitrogen-fixing activity were sterilised and grown in sterilised soil under germ-free conditions. Nitrogen fixation was measured. No acetylene reduction was detectable on the sterile plant, suggesting that microorganisms must be involved. Although no nodules were found on the roots of rice associated with high nitrogen-fixing activity, an anaerobic, rod-shaped, Gram-positive, biotin-requiring bacterium, which formed terminal spores was isolated. This is thought to be a *Clostridium*. It is known, however, that from a single soil sample one can isolate many species of bacterium capable of fixing nitrogen, such as *Desulfovibrio*, *Mycobacterium* and *Beijerinckia*, and it is not yet possible to detect which bacterium is responsible for nitrogen fixation in the rice rhizosphere. A systematic examination of nitrogen-fixing bacteria in the rhizosphere of rice in our experimental field is in progress.

Table 1 Variation in acetylene reducing activity of 50 rice strains

C_2H_2 produced (nmol per plant per h)	No. of strains
0–200	13
200–400	23
400–600	8
600–800	2
800–1,000	3
1,000–1,200	1

Acetylene reducing activity of strains was obtained by averaging the measurements of two plants obtained during the flowering period.

Table 2 Analysis of variance of acetylene reducing activity among 50 rice strains

Source of variation	d.f.	Sum of squares ($\times 10^{-3}$)	Mean square ($\times 10^{-3}$)
Total	99	8,152	
Strain	49	5,576	113.8*
Error	50	2,576	51.5

* Statistically significant at the 1% level.

The significant variation found among rice strains in their degree of association with nitrogen fixation could suggest that the genotype of the plant influences this association. A readily available source of energy is crucial in establishing the association of nitrogen-fixing microbes with rice. If the rice plant excretes available substrates through its root system, microorganisms would take advantage of this to grow and fix nitrogen in conditions which do not repress nitrogenase synthesis. Certain types of plant excrete sufficient carbohydrate to support a sizeable microbial population[3,18]. Not only *Azotobacter* and *Spirillum* benefit, but also strictly anaerobic organisms such as *Clostridium* and facultative organisms[19,20]. The significant correlation between the activity of nitrogen fixation and root weight as well as the genotype of the rice plant could be due to a dependence of nitrogen-fixing microbes on the energy source provided by the plant. As a great deal of energy is needed to support microbial growth and nitrogen fixation, the energy source will be the limiting factor for the plant–bacterium association.

We thank Professor John Postgate for critical comments on the manuscript and Dr K. Shimotohno for advice on gas chromatography and Dr K. Komagata for help with isolation and classification of a nitrogen-fixing bacterium. This study was supported by grant 111905 from the Ministry of Education, Japan.

Y. HIROTA
T. FUJII
Y. SANO
S. IYAMA

National Institute of Genetics,
Yata 1111, Mishima,
Shizuoka-ken 411, Japan

Received 10 July; accepted 3 October 1978.

1. Parker, C. A. *Nature* **179,** 593–594 (1957).
2. Renaudo, G., Balandreau, P. & Dommergues, Y. *Pl. Soil* Spec. Vol. 471–479 (1971).
3. Döbereiner, J., Day, J. M. & Dart, P. J. *J. gen. Microbiol.* **71,** 103–116 (1972).
4. von Bülow, J. F. W. & Döbereiner, J. *Proc. natn. Acad. Sci. U.S.A.* **72,** 2389–2393 (1975).
5. Yoshida, T. & Ancajas, R. R. *Soil Sci. Pl. Nutr.* **16,** 234–237 (1970).
6. Balandreau, J., Miller, C. & Dommergues, Y. *Appl. Microbiol.* **27,** 662–665 (1974).
7. Watanabe, I. *Envir. Role Nitrogen-Fixing Blue-green Algae Asymbiotic Bacteria, Int. Symp.*, Uppsala (1976).
8. Lee, K. K., Castro, T. & Yoshida, T. *Pl. Soil* **48,** 613–619 (1977).
9. Dommergues, Y., Balandreau, J., Rinaudo, G. & Weinhard, P. *Soil Biol. Biochem.* **5,** 83–89 (1973).
10. Burris, R. H., Okon, Y. & Albrecht, S. *Envir. Role Nitrogen Fixing Blue-green Algae Asymbiotic Bacteria, Int. Symp.*, Uppsala (1976).
11. Watanabe, I. *et al. IRRI Res. Pap.* Ser. **3,** 1–16 (1977).
12. Yoshida, T. & Ancajas, R. R. *Soil Sci. Am. Proc.* **37,** 42–46 (1973).
13. Lee, K. K. & Yoshida, T. *Pl. Soil* **46,** 127–134 (1977).
14. Okon, Y., Albrecht, S. L. & Burris, R. H. *Appl. envir. Microbiol.* **33,** 85–88 (1977).
15. Burris, R. H., Okon, Y. & Albrecht, S. L. *Genetic Engineering for Nitrogen Fixation* (ed. Hollaender, A.) 445–450 (Plenum, New York, 1977).
16. Hardy, R. W. F., Holsten, R. D., Jackson, E. K. & Burns, R. C. *Pl. Physiol.* **43,** 1185–1207 (1968).
17. Burris, R. H. *Meth. Enzym.* **24,** 415–431 (1972).
18. Döbereiner, J., Day, J. M. & Dart, P. J. *Pl. Soil* **37,** 191–196 (1972).
19. Willemin, G., Balandreau, J. & Dommergues, Y. *Ann. Microbiol.* **24,** 87–94 (1974).
20. Balandreau, J. *et al. Limitations and Potentials for Biological Nitrogen Fixation in the Tropics* (eds Döbereiner, J. *et al.*) 275–302 (Plenum, New York, 1978).

Identification of a new chicken α-globin structural gene by complementary DNA cloning

THE identification of a new chicken α-globin gene is reported here. Nucleotide sequence analysis of clone pHb1003 shows it to code for an α-subunit with an amino acid sequence different from those normally observed in the adult (the αA and αD). Characterisation of a second clone, pHb1010, carrying the same novel sequence, suggests that this mRNA sequence was present in substantial quantity after very young (2–3-week-old) chickens were treated with phenylhydrazine. The results are therefore similar to those reported previously for sheep and goats, where treatment with phenylhydrazine results in the turn-on of a new stress globin gene. As the ease of obtaining nucleated erythrocytes from avian systems facilitates many types of biochemical studies, we believe that this finding may provide an especially useful system in which to study the molecular events involved in globin gene switching.

Our laboratory has previously reported the complementary DNA (cDNA) cloning of globin mRNA sequences from phenylhydrazine-treated 2–3-week-old white Leghorn chickens[1]. One of these clones, pHb1003, was shown by the ladder DNA sequencing technique of Maxam and Gilbert[2] to contain an α-globin cDNA insert. This determination was based on the discovery of a nucleotide sequence within a pHb1003 *Hae* III insert band which codes for amino acid residues 86–98 of the chicken αA primary structure as determined by Matsuda[3]. These amino acids are highly conserved in most vertebrate α-globin proteins[4]. As such, this sequence is useful for identifying pHb1003 as an α-globin insert, but it does not indicate whether this DNA is derived from the αA or another α gene. Other positions along the α-globin molecule are more variable, most notably positions 102–125, and 46–82 (ref. 4). Restriction mapping data (Fig. 1) enabled us to locate and sequence pHb1003 DNA fragments from the latter variable region. Using the ladder technique, sequences were obtained from a 510-base pair *Alu* I fragment coding for the untranslated region at the 5′ end of the mRNA and amino acids 1–83 of the α globin (Fig. 2). The 69-base pair sequence obtained codes for amino acid residues 60–83 of the α globin. It includes 10 positions where the amino acid sequence is not compatible with the αA sequence and 16 positions which diverge from the αD sequence[5] (Fig. 3). The presence of so many differences argues against the idea that we are observing an allelic form of either the αA or αD genes.

As the αA and αD subunits account for 70% and 30%, respectively, of the α-subunits normally present in adult chicken reticulocytes[6,7], this raises several interesting possibilities. The polypeptide coded by pHb1003 might represent a minor normal

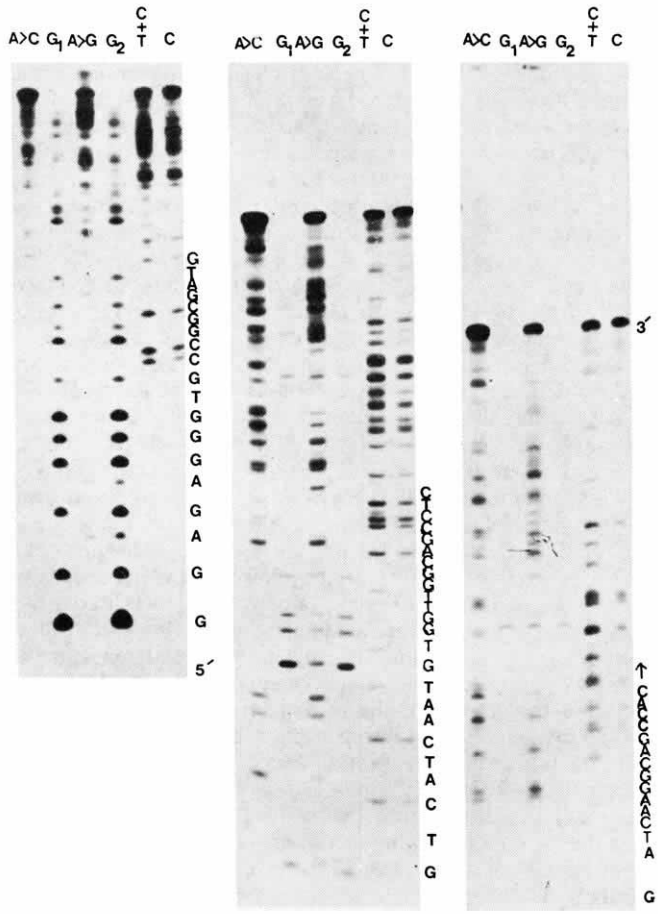


Fig. 2 Autoradiogram of the ladder sequence from a 480-base pair *Alu* I pHb1003 insert band. 100 μg of pHb1003 was digested with *Alu* I and the products were fractionated on a 6% acrylamide 1/28 bisacrylamide preparative gel. The 510-base pair insert fragment was eluted from the gel in high salt overnight. The 5′-terminal phosphates were removed by treatment with bacterial alkaline phosphatase and replaced with the ^{32}P label of λ ^{32}P-ATP using T4 polynucleotide kinase. This DNA, which was labelled at both 5′ termini, was cleaved with *Hae* III, giving two single-end-labelled fragments of 320 and 190 base pairs. The 190-base pair DNA was aliquoted into 5 equal fractions. The aliquots were separately subjected to partial chemical cleavages specific for G, A, A>C, C+T and C. The partial cleavage products were then loaded side by side on 20% acrylamide 1/60 bisacrylamide 7 M urea gels at 3 successive intervals. The radioactive bands were visualised by autoradiography at −70 °C with Kodak Hi-Plus tungstate intensification screens. The nucleotide sequence is read from the 5′ end at the bottom of the gel to the 3′ end at the top. The sequence shown is complementary to the pHb1003 coding strand.

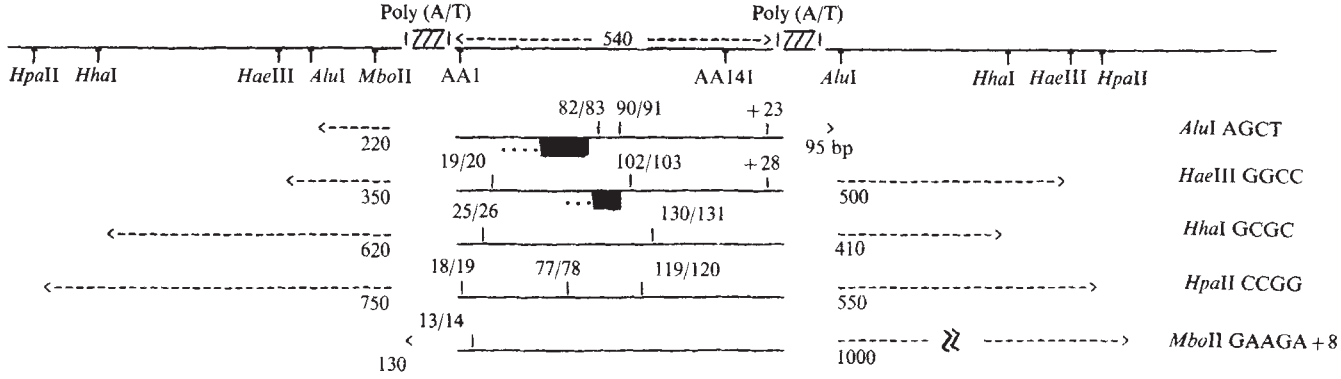


Fig. 1 Restriction map of the pHb1003 α-globin cDNA insert. Insert fragments were identified and sized by electrophoresis of various pHb1003 restriction digests alongside similar cleavages of the pMB9 vector on 6% acrylamide 1/29 bisacrylamide gels. Secondary digests of purified insert cleavages occurring within the 3′ untranslated region are denoted by a '+' followed by the site's distance from the UAA termination codon (in base pairs). Heavy lines are used to indicate the regions which have been sequenced. The exact locations of those restriction sites which have not been sequenced were estimated by extrapolating from those in regions where the nucleotide sequence is known and fitting with the amino acid sequence. The numbers indicate base pairs.

© Macmillan Journals Ltd 1978

```
PHB1003:                                                                                  HpaII                   AluI
coding   5′ AAG AAG GTA GTG GCT GCC TTG ATC GAG GCT GCC AAC CAC ATT GAT GAC ATC GCC GGC ACC CTC TCC AAG CT  3′
comp     3′ TTC TTC CAT CAC CGA CGG AAC TAG CTC CGA CGG TTG GTG TAA CTA CTG TAG CGG CCG TGG GAG AGG TTC GA  5′

           68                  65                  70                  75                  80
α D:       LYS LYS VAL LEU GLY ALA LEU GLY ASN ALA VAL LYS ASN VAL ASP ASN LEU SER GLN ALA MET ALA GLU LEU
           ////////////        ////////        ///                 ///                                 ///
PHB1003:   LYS LYS VAL VAL ALA ALA LEU ILE GLU ALA ALA ASN HIS ILE ASP ASP ILE ALA GLY THR LEU SER LYS LEU
           ////////////        ///             ///         ///     ////////////    ///     ////////////////
α A:       LYS LYS VAL ALA LEU ALA ILE THR ASN ALA ILE GLU HIS ALA ASP ASP ILE SER GLY ALA LEU SER LYS LEU

                           A       T           T                       T   C           A   T       A
α A          A   A       TTG     C       T       C   A   T       T   T   C AGT         TTG AGC   A TTG
codons:    AAG AAG GTX GCX CTX GCX ATA ACX AAC GCX ATA GAC CAC GCX GAC GAC ATA TCX GGX GCX CTX TCX AAG CTX

                       A           A                               A                               A
α D          A   A     TTG         TTG     T           A   T       T   T TTG AGT   A               T TTG
codons:    AAG AAG GTX CTX GGX GCX CTX GGX AAC GCX GTX AAG AAC GTX GAC AAC CTX TCX CAG GCX ATG GCX GAC CTX
```

Fig. 3 Nucleotide sequence from the 5′ terminus of the 510-base pair *Alu*I pHb1003 insert fragment shown in Fig. 1. The corresponding polypeptide sequence is shown juxtaposed to the αA and αB globin primary sequences reported by Matsuda and Takei[3,5]. Sequence homology is indicated by the cross-hatched bars. The amino acids listed as evolutionarily conserved by Dayhoff[4] are enclosed in rectangles. The possibte codons for amino acids 60–83 of the αA and αD are shown below the amino acid comparisons.

adult globin chain or one not previously recognised in 2–3-week-old chicks. Alternatively, it is possible that the stress of phenylhydrazine treatment causes these young chickens to synthesise other globin chains, as is known to be the case in sheep or goats. If so, this stress globin might represent activation of a globin gene not previously identified, or it might represent renewed synthesis of embryonic π or π' α-globin subunits or of haemoglobin H (the α subunit of which is uncharacterised).

The observed variations cannot easily be explained as the result of rearrangements in the cloned DNA; it is unlikely that such rearrangements could cause the observed variations while preserving the correct reading frame, the conserved residues at positions 69, 74, 80 and 83, and the proper spacing between these conserved residues. Nor is it likely that these variations are due to polypeptide sequencing errors: the amino acid sequence in question is bracketed by a tryptic fragment from Lys 61 to Lys 81 and a chymotryptic fragment from Leu 65 to Leu 80. Both these fragments are small and easily sequenced by the Edman (phenylisothiocyanate) and Sanger (dinitrofluorobenzene) methods used by Matsuda *et al.*

To test the hypothesis that pHb1003 might represent only a minor globin sequence, we have carried out a preliminary characterisation of other chicken globin cDNA clones. Twelve globin-insert-carrying plasmids identified from colony hybridisation studies were grown and analysed by cleavage with the restriction enzyme *Hae*III. Of the restriction patterns obtained, one showed a 248-base pair fragment characteristic of pHb1003. All but one of the remaining plasmids contained β-globin inserts as shown by the presence of a 66-base pair *Hae*III fragment. The remaining plasmid, pHb1008, showed a unique pattern and may represent an entirely different α- or β-globin gene sequence (A.Y.L. and Lee, unpublished results). The discovery of a second plasmid, pHb1010, with a restriction pattern similar to that of pHb1003 indicated that this is not a minor sequence in this mRNA preparation. It also made it possible to test directly the possibility of rearrangements in the cloned DNA. The ladder sequencing technique was used to compare the nucleotide sequences of plasmids pHb1003 and pHb1010 in the same region shown in Fig. 3. The sequence of pHb1010 was found to be identical to that found for pHb1003 in this region, but different from the sequence expected for the αA or αD mRNAs, as discussed earlier.

The fact that nine clones out of 12 showed the β-globin restriction pattern suggests two possibilities. The first is that the α-globin sequences were cloned at low frequencies, in which case we might have lost a specific class of α-globin mRNA. The alternative possibility is that α and β sequences were cloned with similar efficiencies but that the α clones give slightly lower signals in the colony hybridisation test so that they are somewhat under-represented among the 13 clones chosen for the most intense signals. Our experimental results show that this latter possibility is correct. When the colony hybridisation test was repeated with probes specific for the α- and β-globin genes (restriction fragments from pHb1001 and pHb1003 were used as probes), approximately equal numbers of α- and β-globin clones were observed (I.W.C., unpublished results).

On the basis of these results, we favour the hypothesis that the stress of phenylhydrazine treatment in these young chickens may cause the turn-on of synthesis of a stress globin. The complete sequence determination of the pHb1003 clone, which is now underway, will be helpful in establishing whether it codes for the embryonic π or π' α-globin subunits, or the α subunit of haemoglobin H, or for a novel stress globin such as occurs in sheep or goats. Interestingly, the amino acid variations observed between positions 60 and 83 generally retain the polarity and charge observed in the corresponding region of the αA polypeptide sequenced by Matsuda *et al.* With the exception of a Thr-to-Ile shift at amino acid residue 67, all the replacements combine to prevent an overall change in polarity or charge. The Asn-to-Glu change at residue 68 is opposed by a Glu-to-Asn change at position 71. The Ser-to-Ala shift at amino acid 77 is compensated by an Ala-to-Thr change at position 79. The remaining changes at residues 63, 64, 70 and 73 all involve uncharged nonpolar amino acids.

Amino acids 60–83 occur in the EF 'corner' region of the α-globin polypeptide between the E and F helices (residues 52–71 and 80–88, respectively, by Kendrew's nomenclature[9,10]). The EF region bends around the haem and forms a part of the hydrophobic centre of the intact haemoglobin molecule. This suggests that for the chicken α-globin polypeptide the maintenance of correct charge and polarity in such bend regions may be more important than other features of protein stereochemistry.

This work was supported by USPHS grants GM 18586 and HL 21831. A.L. was supported by NCI predoctoral training grant 4–783825–32923–7.

Note added in proof: One of us (A.Y.L.) has completed the DNA sequence of pHb1003 (manuscript in preparation). The polypeptide sequence encoded by this plasmid insert differs from the adult αA globin at 23 positions and from the αD at 63 positions. At the same time the predicted tryptic cleavage products of this sequence would differ from those reported for the π and π' embryonic α-type globins. We are therefore led to conclude that pHb1003 encodes a previously uncharacterised αA-like globin.

IAN W. CUMMINGS
ALVIN Y. LIU
WINSTON A. SALSER

Department of Molecular Biology,
University of California at Los Angeles,
Los Angeles, California 90024

Received 28 August; accepted 28 September 1978.

1. Padayatty, J. *et al. Nucleic Acids Res.* (submitted).
2. Maxam, A. & Gilbert, W. *Proc. nam. Acad. Sci. U.S.A.* **74,** 570 (1977).
3. Matsuda, G., Takei, H., Wu, K. & Shiozawa, T. *Int. J. Protein. Res.* **111,** 173 (1971).
4. Dayhoff, M. *Atlas of Protein Sequence and Structure,* Vol. 5 (National Biomedical Research Foundations, Silver Spring, 1972).
5. Takei, H., Ota, Y., Wu, K., Kiyohara, T. & Matsuda, G. *J. Biochem.* **77,** 1345 (1975).
6. Schalekamp, M., van Goor, D. & Sungerland, R. *J. Embryol. exp. Morph.* **28,** 681 (1972).
7. Moss, B. & Thompson, E. O. P. *Aust. J. biol. Sci.* **22,** 1455 (1969).
8. Brown, J. & Ingram, V. *J. biol. Chem.* **249,** 3960 (1974).
9. Kendrew, J. *et al. Nature* **185,** 422 (1960).
10. Perutz, M. *Br. med. Bull.* **32,** 195 (1976).

matters arising

Seismic evidence for Mesozoic sedimentary troughs on the Hebridean continental margin

JONES'S interpretation of seismic refraction lines on the NW British continental shelf[1] provides welcome corroboration of the Mesozoic sedimentary basin between the Flannan Isles and Lewis first postulated by Bullerwell (see refs 2–4). As we have already suggested that the Rockall Trough developed by seafloor spreading in the early Permian[5–7], we wish we could also welcome Jones' interpretation of the structure of the NW British margin, which suggests that the development of the Trough, at least in part, goes back to Permo-Triassic times. However, we interpret the high velocity refractor ($V_P = 4.4 \pm 0.3$ km s^{-1}) as corresponding merely to the top of a layer of Palaeocene–Eocene basalts, because (1) regional mapping and study of interval velocities using commercial 24-fold reflection lines and aeromagnetic coverage shows that basalts extend southwards from the Faeroes over almost the whole of the Rockall Trough north of Anton Dohrn seamount (as was suggested by Roberts[8]) and also over the outer NW Scottish shelf from the Wyville–Thomson Ridge to about 58° N; (2) the high-frequency negative magnetic anomaly field characteristic of these basalts at shallow depth is clearly seen both on Jones' magnetic profile[1] and on the IGS aeromagnetic map[2]; and (3) the refractor velocity is similar to that for basalts elsewhere on the NW European margin[9–11], but probably too high for Permo–Triassic sediments; although the interval velocity for pure Permo–Triassic sandstones at ~4 km depth in wells in the Viking Graben[12] and West Shetland Basin[13] can exceptionally attain values as high as 4.5 km s^{-1}, the usual range of velocity is 3.6 – 4.1 km s^{-1} for Permo–Triassic sandstones in these areas. The Permo-Triassic beneath the Little Minch has a $V_P \sim 3.9$ km s^{-1}, obtained by refraction shooting[14]. Above the basalts, the refractors with apparent $V_P \sim 2.8$–3.1 km s^{-1} could be due to Quaternary glacial till; measurements on data from the West Shetland area show that glacial till up to 100 m thick can have interval velocities $V_{int} \sim 2.6$–2.9 km s^{-1}. However, the Tertiary sediments between the Quaternary and the basalts have a lower V_{int}, so the top of the basalts is somewhat shallower than that given by Jones' profiles.

The gravity suggests that there may be sediments beneath the basalts, but we have no way of estimating their age. Thus we conclude that Jones' results have no relevance to the early evolution of the Rockall Trough, and, furthermore, his attempt to estimate the hydrocarbon prospects of this part of the margin is misleading; prospects here are poorer than elsewhere, because of the basalt cover. However, our reconstruction[7] suggests that in mid-Jurassic times Orphan Knoll was situated at the south-east margin of the Rockall Trough. The neritic Bajocian sands found beneath the Knoll[15] may therefore occur elsewhere along the margins of the Trough, and may also overlie downfaulted basins of Carboniferous age on the outer continental margin, developed during an intra-continental rifting phase[6] before early Permian spreading. Therefore along the margins clear of Tertiary basalt cover, for example, the south-east Faeroe–Shetland Trough margin between 60 and 62° N, and both margins of the Rockall Trough south of about 58° N, the hydrocarbon prospects may be very good.

D. K. SMYTHE
N. KENOLTY
Institute of Geological Sciences, Marine Geophysics Unit, Murchison House, West Mains Road, Edinburgh, UK

M. J. RUSSELL
Department of Applied Geology, University of Strathclyde, James Weir Building, 75 Montrose Street, Glasgow, UK

1. Jones, E. J. W. *Nature* **272**, 789–792 (1978).
2. Bullerwell, W. *Aeromagnetic Map of Part of Great Britain and Northern Ireland, Sheet 12* (Geol. Surv. G.B., 1968).
3. Eden, R. A., Wright, J. E. & Bullerwell, W. *Rep. Inst. geol. Sci.* No. 70/14, 111–128 (1971).
4. Dunham, K. *The Sub-Pleistocene Geology of the British Isles and Adjacent Continental Shelf* (Inst. geol. Sci., 1972).
5. Russell, M. J. in *Implications of Continental Drift to the Earth Sciences* (eds Tarling, D. H. & Runcorn, S. K.) 581–597 (Academic, London, 1973).
6. Russell, M. J. *Scott. J. Geol.* **12**, 315–323 (1976).
7. Russell, M. J. & Smythe, D. K. in *Petrology and Geochemistry of Continental Rifts* (eds Neumann, E. R. & Ramberg, I. B.) 173–179 (Reidel, Dordrecht, 1978).
8. Roberts, D. G. *Deep-Sea Res.* **18**, 353–360 (1971).
9. Pálmason, G. *Tectonophysics.* **2**, 475–482 (1965).
10. Chalmers, J. A., Dobinson, A., Mould, A. & Smythe, D. K. *Geophys. J. R. astr. Soc.* **49**, 288 (1977).
11. Talwani, M. & Eldholm, O. *Bull. geol. Soc. Am.* **83**, 3575–3606 (1972).
12. Kent, P. E. *J. geol. Soc. Lond.* **131**, 435–468 (1975).
13. Cashion, W. W. in *Offshore Europe 75*, Paper OE-75-216 (Spearhead, Kingston-upon-Thames, 1975).
14. Smythe, D. K., Sowerbutts, W. T. C., Bacon, M. & McQuillin, R. *Nature phys. Sci.* **236**, 87–89 (1972).
15. Laughton, A. S. *et al.* in *Init. Rep. DSDP* Leg 12 33–159 (1972).

JONES REPLIES—The uncertainties involved in estimating stratigraphical ages from seismic velocities in the Hebridean region were emphasised in my paper[1] and I am therefore grateful to Smythe *et al.* for their early comments. Although these are primarily based on their unpublished reflection profiles, velocity data and regional maps, some further pertinent remarks can be made.

Smythe *et al.* accept my conclusion that Mesozoic sediments occur immediately west of Lewis, an area in which the age of the sediment cover was previously undefined[2]. My inference that Mesozoic sediments also lie close to the sea floor north-west of the Flannan Ridge is based on the observation that the 2.8–3.1 km s^{-1} values on lines JM-9 and JM-10B are significantly higher than published velocities in Tertiary–Quaternary sequences[1]. Thus constrained, I was unable to interpret the deeper 4.0–4.4 km s^{-1} refractor as the top of a pile of Tertiary basaltic lavas. Using 'measurements on data' from an unspecified location near the Shetlands, Smythe *et al.* attribute the 2.8–3.1 km s^{-1} values to glacial till, which then allows them to suggest that the 4.0–4.4 km s^{-1} layer is much younger than I proposed. As they do not indicate the water depths in which the high Pleistocene velocities were determined, the method of measurement and also local velocity changes, it is not clear whether a direct comparison with shallow refractor velocities near the Flannan Isles is justified. Are their 2.6–2.9 km s^{-1} values sufficiently typical of the Pleistocene to permit an extrapolation over a large area? Their results may be of regional importance but as careful mapping by Eden[3] and others has revealed marked and often rapid lateral variations in Pleistocene deposits off northern Britain, Smythe *et al.* need to provide additional evidence to show that thick, high velocity boulder clay is likely to lie near the shelf break, some 300 km south-west of the Shetlands. They must also account in their interpretation for the presence of pre-Quaternary reflectors[4,5] within the westwards continuation of their proposed Quaternary sequence.

In their discussion of the 4.0–4.4 km s^{-1} refractor, four references are misleadingly quoted (their refs 10–13) to add weight to their contention that it represents Tertiary basaltic lavas and not pre-Jurassic sediments. Seismic velocities for basalts are omitted in ref. 10. Talwani and Eldholm[11] ascribe velocities near 4.4 km s^{-1} on the Norwegian margin, not to basalts, but to pre-Cretaceous sediments. A limiting 4.5 km s^{-1} for Permo–Triassic sediments is not included in their refs 12, 13. My values of 4.0–4.4 km s^{-1} seem to be well

© Macmillan Journals Ltd 1978

within the permissible range for Permo-Trias as the refraction profiles of Browitt[6] show velocities of 4.6–4.8 km s^{-1} at a level in the West Shetland Basin where Cashion's sections[7] reveal thick sediments of this age. Thus, pre-Jurassic deposits cannot be dismissed at the depths I have indicated. They correctly note that magnetic anomalies occur over the 4.0–4.4 km s^{-1} layer but, rather than arising from Tertiary lavas, these can be interpreted as reflecting small basic intrusions within a thick Mesozoic section and magnetisation contrasts in the Precambrian basement.

In denying that my results are relevant to the early evolution of the Rockall Trough, Smythe *et al.* have ignored my demonstration of major faulting and crustal subsidence. As they accept a Mesozoic age for the Flannan Trough they must also accept the existence on the continental margin of NNE-trending faults which were active during Mesozoic time. As these run approximately parallel to the continental slope north of St Kilda and have increasingly larger displacements northwestwards[1], it is difficult to resist the conclusion that the faulting is directly associated with the major phase of downwarping in the Rockall Trough. My section showing thick pre-Jurassic sediments north-west of the fault-bounded Flannan Ridge is compatible with the Permian opening first suggested by Heirtzler and Hayes[8] but it does not establish that seafloor spreading took place at this time.

Commenting on petroleum prospects off northern Britain, Smythe *et al.* offer a comparison with other parts of the Rockall Trough that hinges on unpublished data and a geometrical argument which is questionable because the extent of oceanic crust in the region is unknown. Several parts of the Rockall Trough clearly deserve detailed commercial exploration, although my appraisal of prospects would be less sweeping than that given by Smythe *et al.* Moreover, in view of the geophysical indications of thick sediments and major Mesozoic faulting west of Hebrides, it seems premature for them to take a firm, pessimistic stand on a large section of the Scottish margin, especially since drilling investigations are at such an early stage.

E. J. W. JONES
Department of Geology,
University College,
Gower St, London WC1, UK

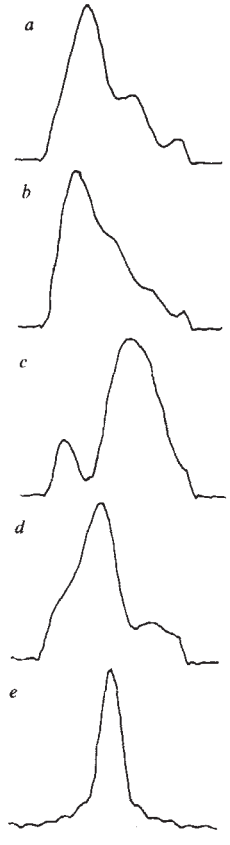
1. Jones, E. J. W. *Nature* **272**, 789–792 (1978).
2. Dunham, K. *The Sub-Pleistocene Geology of the British Isles and the Adjacent Continental Shelf* (Inst. geol. Sci., 1972).
3. Eden, R. A., Holmes, R. & Fannin, N. G. T. *Rep. Inst. geol. Sci.* No. 77/15 (1978).
4. Jones, E. J. W., Ewing, M., Ewing, J. I. & Eittreim, S. L. *J. geophys. Res.* **75**, 1655–1680 (1970).
5. Ruddiman, W. F. *Geol. Soc. Am. Bull.* **83**, 2039–2062 (1972).
6. Browitt, C. W. A. *Nature* **236**, 161–163 (1972).
7. Cashion, W. W. in *Offshore Europe 75*, Paper OE-75-216 (Spearhead, Kingston-upon-Thames, 1975).
8. Heirtzler, J. R. & Hayes, D. E. *Science* **157**, 185–187 (1967).

Object reconstruction from intensity data

GULL AND DANIELL[1] have recently proposed a new maximum entropy algorithm for the reconstruction of an object field from incomplete and noisy measurements of its Fourier transform. This new algorithm may encourage other workers to apply maximum entropy techniques to their own problems. For this reason we point out here that the application of maximum entropy techniques for object reconstruction in the absence of phase information could lead to wholly false conclusions due to ambiguities. We shall illustrate some of the problems using a specific example. Many such examples may easily be generated and the one chosen here is not unusual.

Figure 1 shows four different, real and positive objects which all give rise to the scattered intensity shown in Fig. 1*e*. Families of dissimilar objects whose Fourier transforms have the same modulus occur whenever the scattered intensity has non-zero minima[2]. Such minima indicate the presence of complex zeros in the scattered field. For each such zero there are two possible positions, which correspond to different objects whose Fourier transforms have identical moduli. The possible positions for these zeros are symmetrical with respect to the real axis and thus one speaks of 'flipping' zeros about this axis in order to generate different objects. The example in Fig. 1 was produced by 'flipping' such zeros. Clearly the four objects thus created are very different, for example, in the number and position of their peaks.

Fig. 1 All four different real positive objects shown in *a*, *b*, *c* and *d* produce the same far field intensity, shown in *e*. The four different objects have been generated by reflecting or 'flipping' the zeros of $F(z)$ about the x-axis, as described in the text.



The maximum entropy algorithm should converge to one of these possible solutions but the only criterion it has as a basis for the selection of that solution is that of smoothness, as all the possibilities are real and positive. However, inspection of the four possible solutions shows that none may be disregarded as unphysical—they are all 'smooth'. It would only be possible to identify one of these possibilities as the solution if one has some extra information, such as approximate phase values which may provide a basis for such discrimination. On the basis of the intensity alone there can be no unique solution, unless all minima are zero. The selection of one solution using an entropy criterion is dangerous. For example, if the algorithm chose the third solution one might conclude that the object consisted of two distinct peaks, whereas the other possible solutions show that such a conclusion is unfounded.

The example given here is simpler than may be expected in general—there are few complex zeros—and thus 'real' data may be expected to result in even greater ambiguity. In these conditions it is vital that some measure of the ambiguity is established. The maximum entropy algorithm does not give such a measure[3].

For convenience the analysis used to produce the example in Fig. 1 is one-dimensional but similar ambiguities will exist in two-dimensional systems.

M. A. FIDDY*
A. H. GREENAWAY†
Physics Department,
Queen Elizabeth College,
Campden Hill Road,
Kensington, W8, UK

*Present address: Dept of Electronic and Electrical Engineering, University College London, Torrington Place, London WC1, UK.

†Present address: Dépt d'Astrophysique, Université de Nice, Parc Valrose, 06034 Nice-Cedéx, France.

1. Gull, S. F. & Daniell, G. J. *Nature* **272**, 686 (1978).
2. Burge, R. E., Fiddy, M. A., Greenaway, A. H. & Ross, G. *Proc. R. Soc.* **A350**, 191 (1976).
3. Dainty, J. C., Fiddy, M. A. & Greenaway, A. H. *IAU/URSI Colloquia on Formation of Images from Spatial Coherence functions in Astronomy*, No. 49 (Groningen, The Netherlands, in the press).

Human reproduction reconsidered

MAY'S article in News and Views, '*Human Reproduction Reconsidered*'[1], fails to mention that the research on the demography of the !Kung hunter-gatherers, reviewed by Short and by

0028–0836/78/0276—0421$01.00
© Macmillan Journals Ltd 1978

Kolata, is from my work. A brief report can be found in *Kalahari Hunter-gatherers*[2]. A full account, including data, methods and analyses (including a comparison with the Hutterites and a consideration of the implications of the Frisch hypothesis) will be found in my forthcoming book[3].

The figure of 15.5 yr for the mean age at menarche for !Kung girls (quoted by Kolata from a telephone interview) has been superceded by a mean of 16.5 yr, calculated from a larger group of observations. !Kung women may be anovulatory during some substantial portion of their birth intervals (mean length 35.4 months for all intervals observed during an eleven year period, $n = 102$) but they are not amenorrheic for this whole period.

It is tempting but premature to generalise from the !Kung to hunter-gatherers in general and hence to prehistoric populations. The low fertility of the !Kung suggests a stimulating hypothesis on the nature of population dynamics in such groups, as discussed by May, but exploring that hypothesis will require considerable additional research on other groups.

NANCY HOWELL
Department of Sociology,
University of Toronto,
Toronto, Ontario, Canada

1. May, R. M. *Nature* **272**, 491–495 (1978).
2. Howell, N. in *Kalahari Hunter-gatherers* (eds Lee, R. & DeVore, I.) (Harvard University Press, Cambridge, 1976).
3. Howell, N. *Demography of the Dobe !Kung* (Academic, New York, in the press).

The neural representation of visual space

DRASDO[1], and Lennie[2], attribute Rolls and Cowey[3] with the demonstration that monkey cortical magnification factor, M, varies as the square root of the ganglion cell density which deals with the corresponding region of visual field. Lennie's discussion is based on the development of this relationship for monkey and man. In fact, Rolls and Cowey report M to be proportional to ganglion cell density deg^2, not to its square root, both within and outside the foveal area.

Such proportionality between retinal ganglion cell density deg^2 and M has also been described at several stations of the visual pathways in a variety of species[4–7]. It has been a long-standing puzzle as to why a linear measurement, M, should be proportional to a retinal cell density.

The outcome of Drasdo's data manipulations, that human cortical magnification factor is proportional to the square root of the ganglion cells serving the corresponding region of visual field (that is, of his receptive field density, D_r), is thus not at all in agreement with the findings of Rolls and Cowey. If valid, it is the first instance of such a relationship and implies that the human visual cortex is organised in a quite different fashion from that of saimiri or macaque, or indeed from that of cat and rabbit.

However, the data on which Drasdo's conclusions are based are not sound enough to warrant a belief that the human cortex deviates so markedly from the organisation apparently common to a variety of species. The ganglion cell distributions which he pooled[8–10] were all obtained by the unreliable technique of counting on radial sections not subject to an Abercrombie correction[11]. Also, there is no evidence of common criteria for ganglion cells, shrinkages were not measured and the distributions plotted separately can support quite different interpretations from those of Drasdo. It perhaps suffices to point out that the results of Van Buren indicate densities of from one to more than two orders of magnitude greater than those of Oppel at the equivalent eccentricity. Correction for optical magnification is a nicety in such circumstances.

Of course, it would be pleasing if Drasdo's conclusions were valid, because the recent demonstration[12] of proportionality between human acuity and cortical magnification factor could then relate acuity directly to ganglion cell angular separation. In my opinion the evidence is against such a relationship in man[13] or cat[14] but, as Lennie points out, knowledge of the distribution of the specific ganglion cell class involved in resolution tasks may elucidate the matter. *Note added in proof:* Lennie's editorial remains explicitly misleading. This might have been avoided had Drasdo emphasised that he claimed differences between man and monkey which necessitated analysis in terms of $\sqrt{D_r}$, rather than D_r, and M. However, I remain unconvinced; the densities of Van Buren and Oppel differ more than 100-fold at the same eccentricity in the original data quite independent of errors introduced by replotting; similar large discrepancies occur throughout their data and preclude pooling.

A. HUGHES
Department of Physiology,
John Curtin School of Medical Research,
Australian National University,
PO Box 334, Canberra City,
ACT 2601 Australia

1. Drasdo, N. *Nature* **266**, 554–556 (1977).
2. Lennie, P. *Nature* **266**, 496 (1977).
3. Rolls, E. T. & Cowey, A. *Expl Brain Res.* **10**, 298–310 (1970).
4. Jacobsen, M. *Q. J. exp. Physiol.* **47**, 170–178 (1962).
5. Sanderson, K. *J. exp. Brain Res.* **13**, 159–177 (1971).
6. Hughes, A. *Documenta ophth.* **30**, 33–159 (1971).
7. Webb, S. V. & Kaas, J. H. *Vision Res.* **16**, 1247–1254 (1976).
8. Vilter, V. *C. r. Séanc. Soc. Biol.* **148**, 220–223 (1954).
9. Van Buren, J. M. *The Retinal Ganglion Cell Layer* (Thomas, Springfield, 1963).
10. Oppel, O. *Albrecht v. Graefes Arch. Ophthal.* **172**, 1–22 (1967).
11. Abercrombie, M. *Anat. Rec.* **94**, 239–247 (1946).
12. Cowey, A. & Rolls, E. T. *Expl Brain Res.* **21**, 447–454 (1974).
13. Hughes, A. *Handbook of Sensory Physiology*, Vol. 7/5 A (Springer, Berlin, 1977).
14. Tusa, R. J., Palmer, L. A. & Rosenquist, A. C. *J. comp. Neurol.* **177**, 213–235 (1978).

DRASDO REPLIES—Some responsibility is accepted for the scepticism evinced by Hughes, whose comments on my letter[1] can be traced to two irregularities on points of detail[2]. However, this scepticism should not extend to the material contents of the letter which can be shown to be unaffected. As Hughes indicates, the root sign in the first paragraph is regrettably misplaced. This relationship observed in animals was stated correctly in the following paragraph.

Apparent differences between the vision of man and other animals should not cause surprise. Even in other primates, predation, falls and fighting contribute largely to mortality[3]; survival may be related to peripheral and tectal vision, and correspondingly the human extra foveal retina has fewer presumed transient and feature-detecting ganglion cells[4].

The ganglion cell density distributions differed systematically; literally, to pool them would have been reprehensible. Instead, a mean line was taken between the two sets for each meridian (see Fig. 1, ref. 1). Shrinkage corrections for data of Oppel and Van Buren were based on fovea–papilla distance, which further reduces disparities by correcting for eye size. Paradoxically, Oppel's[5] original estimated total count fell short of Van Buren's[6] by only 40%, despite the extreme discrepancies referred to by Hughes.

I am indebted to Hughes, who inadvertently brought to my attention an incorrectly plotted point which, when corrected[2], reduced the discrepancy between cells and receptive fields in the area affected by excavation to 3%, giving further support to the calculation (see legend to Fig. 1, ref. 1).

The possibility of an overcount of ganglion cells is well known[6,7], but as with the other data used these were the best estimates available and all show an impressive alignment with the empirical equation which remains attractive for its simplicity and usefulness in the present state of knowledge.

N. DRASDO
Department of Ophthalmic Optics,
University of Aston in Birmingham,
Birmingham, UK

1. Drasdo, N. *Nature* **266**, 554–556 (1977).
2. Corrigendum *Nature* **272**, 840 (1977).
3. Schultze, A. H. *The Life of the Primates*, 130, 193–199 (Wiedenfield and Nicholson, London, 1969).
4. Dubin, M. W. in *The Eye*, Vol. 6, eds Davson, H. & Graham, L. T. J. 249–254 (Academic, London, 1974).
5. Oppel, O. *Albrecht v. Graefes Arch. Ophthal.* **172**, 1–22 (1967).
6. Van Buren, J. M. *The Retinal Ganglion Cell Layer*, 128–130 (Thomas, Springfield, 1963).
7. Rolls, E. T. & Cowey, A. *Expl Brain Res.* **10**, 298–310 (1970).

0028–0836/78/0276—0422$01.00
© Macmillan Journals Ltd 1978

reviews

Chinese technical endeavour

N. Sivin

The Shorter Science and Civilisation in China. Vol. 1. An abridgement by C. A. Ronan of J. Needham's original text. Pp.326. (Cambridge University Press: London, New York and Melbourne, 1978.) £7.95.

FOR twenty-four years, in eight volumes of *Science and Civilisation in China* (still unfinished), several associated books, and a multitude of essays, Joseph Needham and his collaborators have been demonstrating the richness and importance of China's scientific and technical traditions before modern times. They have surveyed historical studies by hundreds of scholars in China, Japan and the West, supplemented them with explorations of their own, and woven them into a coherent whole for the first time. The assumption that Europe produced the only civilisation ever capable of systematically understanding and manipulating nature has been current too long to disappear merely because it has been massively disproven. But now that notion is a badge of wilful ignorance. To replace it Needham has offered his ecumenical vision of man's relationship to nature as the one area of experience in which the global effort has been common and convergent over the past several thousand years.

Because the intent of his work is "extensive and pioneering", Needham's books—even collections of essays and lectures such as *The Grand Titration* and *Clerks and Craftsmen in China and the West*—have been speculative in the best sense, closely argued, and massively documented. The latest volume of *Science and Civilisation in China*, a historical survey of alchemy, for instance, contains, alongside 262 pages of text, 169 pages of closely spaced bibliography. It brings the number of illustrations in the series so far to 1373.

After using Needham's publications in courses on Chinese science and medicine at MIT and the University of Pennsylvania for a dozen years, I can report that bright undergraduates with some experience in and taste for historical studies usually find these books readable and stimulating; that other students (especially science and engineering students, whose ability to learn outside their technical fields tends to atrophy in American universities) find them too dense; and that they are too expensive for students to own. The demand for a cheap, simplified version of Needham's survey, for students and others who want to live with the issues he raises, is therefore obvious. This book will satisfy it admirably.

The volume in hand covers Volumes I and II, on history, geography, the travel of scientific knowledge, and the history of philosophy, all seen from the special vantage point of science. Practically every aspect of Chinese technical endeavour since the New Stone Age is touched upon, with constant attention to corresponding aspects of other civilisations. This condensed version begins a new series of unspecified length.

Colin Ronan (who refers to Needham in the third person throughout) is an experienced author of books on science and its history for a general readership. He provides a straightforward condensation, with well-conceived minor rearrangements. He expresses himself clearly and simply, and preserves a good deal of Needham's nuance. The new version provides slightly over a third as much text, nearly as many tables, and more than half as many illustrations (some new), as the volumes abridged. The book is reasonably priced by current standards, and no doubt will become the first of Needham's books on China (in English, at least) to be published in paper covers.

The idea of a multi-volume abridgement is not intuitively obvious, but there is a great deal to be said for a level of detail that will put the gist of Needham on physics and physical technology, or on alchemy, protochemistry, and chemical technology, into one volume. Since the final fascicle of *Science and Civilisation in China* is nowhere in sight, the alternative to a serial abridgement would be an indefinite wait for a superficial single volume.

This book has several limitations of which its readers should be aware. Most important, it is based on writing roughly a quarter-century old. Needham's intellectual reach and the open-endedness of his arguments have made his writing less prone to obsolescence than most other Sinological surveys of the 1950s. His is still the only one that emphasises the scientific and technological components. Nevertheless, our understanding of many matters has been altered fundamentally by the burst of scholarship on China generally, and Chinese science particularly, in the past two decades.

Ronan has interpolated concise notes on important new discoveries and changes in chronology, and has corrected minor errors. But major re-writing, which would be the only way to incorporate important re-interpretations by Needham and many other scholars, is clearly not feasible. None of the investigators capable of fundamental re-writing — which would amount to a new synthesis of the field —is prepared to undertake it, and none known to me could produce a book as readable.

I find Ronan's condensation generally faithful. There are a few points where Needham's assertions are mashed beyond recognition—for example, to produce assertions that the Taoists favoured scientific logic (p83), that Shen Kua's late-eleventh-century account (publication misdated, p157) "is probably the first record in any language of the precipitation of metallic copper by iron", and that ancient Tantric Buddhists "adopted in their symbolism what we might call 'electrical energy'" (p271).

The bibliography is short, but not too short. It would have been more serviceable if the many mediocre and outmoded books had been omitted and more references to important introductory works and critical appraisals published in the past decade or two had been substituted for them.

Ronan's own view of the Chinese tradition, as set out in his brief introduction, puzzles me. To give only one example, the assertion that "before the Jesuit arrival Chinese science was quasi-empirical—based on observation and experience, with theory in a comparatively undeveloped state" ignores not only important parts of *Science and Civilisation in China* but many studies published by other historians in recent decades.

This and other puzzling notions in the introduction do not intrude upon the summary of Needham's arguments,

but an introduction that explained the most important changes in our understanding in the past twenty years or so would have been a great help to readers.

All in all this first volume of *The Shorter Science and Civilisation in China* provides a splendid orientation to the cultural matrix of science and technology. It will motivate many readers to seek out more recent discoveries and interpretations. A large public will await Ronan's coming adaptations of Needham's histories of science, techniques and medicine. □

N. Sivin is Professor of Chinese Culture and of the History of Science at the University of Pennsylvania.

Electrical breakdown phenomena in gases

Electrical Breakdown of Gases. Edited by J. M. Meek and J. D. Graggs. Pp. 878. (Wiley: Chichester, UK, and New York, 1978.) £35; $75

THE book of the same title published in 1953 is well known as a standard reference book to engineers and physicists concerned with the many aspects of electrical breakdown phenomena in gases. This welcome new version, edited by the original authors, updates much of the material in the original volume and takes account of the many advances made over the past 25 years. Inevitably the contents reflect changes of emphasis and new areas of current interest. Discussion of lightning discharges and glow-to-arc transitions is now omitted but a chapter on laser induced breakdown has been included.

As in the original volume, chapter 1 is concerned with a discussion of the fundamental collision processes relevant to breakdown phenomena. This includes basic definitions and a description of electron scattering processes, negative ion formation, collisions involving molecular species, electron–ion and ion–ion recombination, and drift and diffusion of electrons and ions.

Chapter 2 is concerned with vacuum breakdown including a discussion of field emission, microdischarges and microparticle pre-breakdown phenomena. The discussion of spark breakdown in uniform fields in chapter 3 includes a full description of pre-breakdown ionisation in gases in steady-state conditions, electron avalanches, ionisation coefficients, and temporal growth of ionisation in impulse conditions.

A general treatment of corona discharges is given in chapter 4, and chapter 5 is concerned with spark breakdown in non-uniform fields. Breakdown voltage characteristics in both uniform and non-uniform fields are considered in chapter 6 for a wide range of gases and gas mixtures.

Chapter 7 deals with the time lag between the application of an impulse voltage and the consequent breakdown of a gas in terms of the so-called 'statistical time lag' and the formative time lags associated with the initiation and development of a discharge across a gap. In chapter 8, high frequency breakdown of gases is discussed in terms of radio and ultra high frequencies in air and sulphur hexafluoride.

Laser induced electrical breakdown of gases is discussed in chapter 9. The intensity of a focused monochromatic beam is considered, together with its application to multiphoton absorption, ionisation and breakdown. Spark channels, which in recent years have become important in the design of opto-electronic devices of high sensitivity are considered in chapter 10. A short description of electrode effects is given in chapter 11.

Nine different well established researchers have contributed to this volume and they and the editors should be congratulated in achieving an unusual uniformity of style for a book of this type. Unnecessary overlap of material between chapters is minimal and the many excellent diagrams and illustrations throughout the book add to its attractiveness. It is a very readable book and should appeal to specialist researchers, postgraduate students, and readers whose interests lie only on the periphery of the general field of breakdown phenomena.

H. B. Gilbody

H. B. Gilbody is Professor and Head of the Department of Pure and Applied Physics at The Queen's University of Belfast, Northern Ireland.

Liquid properties

Liquids and Their Properties: A Molecular and Macroscopic Treatise with Applications. By H. N. V. Temperley and D. H. Trevena. Pp.274. (Wiley: Chichester, UK and New York, 1978.) £15.

THE research work of the two authors of this book, Professor Temperley and Dr Trevena, lies, respectively, in the statistical mechanics of phase transitions and in the experimental study of the physical properties of liquids. These fields exemplify the 'molecular' and the 'macroscopic' of the book's subtitle.

In the first field we have chapters 1–4 on intermolecular forces and statistical mechanics, chapter 10 on liquid structure and chapter 12 on transport properties. The rest of the book is mainly macroscopic: changes of phase, hydrodynamics and acoustics, liquids under pressure and tension (an interest shared by the authors), the gas-liquid surface, mixtures and solutions, and non-newtonian liquids. There is a final chapter on the unusual properties of liquid helium.

To cover this range of material in 270 pages allows only an elementary treatment. The book aims to provide "what every good scientist, whether in research, teaching, industry or medicine should know about liquids." It is, perhaps, most successful in doing this on the macroscopic side, where the more descriptive narrative places less demands on the reader. The classical equations of hydrodynamics are derived painlessly with scarcely a result in vector notation that is not followed by its cartesian equivalent. Statistical mechanics is less easily handled this way and its equations have to be stated rather than derived if the book is to cover the wide range of subjects promised by the chapter titles.

Within these limitations the authors succeed in giving convincing and easily read accounts of their chosen topics. They are not entirely successful in the admittedly difficult job of relating the molecular side to the macroscopic. Moreover, some of the molecular theories discussed are now quite obsolete, such as the 'hole' and 'tunnel' theories of liquids and the lattice theories of liquid mixtures. There is a place in books of this kind for theories that are technically obsolete if they are physically informative (the authors rightly describe the theories of van der Waals), but not for theories which are now obsolete and which never did do justice to the basic physics. More recent and useful theories are, however, also described clearly, with particular emphasis on those put forward by the late Professor J. G. Kirkwood, to whom the authors rightly pay tribute.

The book is well produced and unusually free from misprints.

J. S. Rowlinson

J. S. Rowlinson is Professor of Physical Chemistry in the University of Oxford, UK.

Cosmic travelogue

Beyond the Moon. By P. Maffei. Pp. 376. (MIT Press: Cambridge, Massachusetts, and London, 1978.) $12.50; £8.75.

Beyond the Moon is a cosmic travelogue that takes us at ever-increasing velocities from the inner Solar System to the depths of the Universe. The purpose of adopting this unusual approach is to heighten the reader's awareness of the vastness of space. Some feel for the immensity of the heavens is conveyed by occasionally looking back at the Solar System until the Earth and Sun disappear from view. To adopt such a style of writing is a bold stroke: on the one hand it may captivate the curious reader, but there are points at which the description of the journey becomes altogether too precious and perhaps out of place in a scientific book.

Thus, on p167 we plunge into the cliché of holiday guide "Having witnessed the phastasmagorical spectacles of double stars and the cataclysms of novae and supernovae, we now set our sights on other goals . . .". Just round the corner to the Orion nebula: ". . . we tend at first to hold our breath, as if for fear that these tenuous veils will come undone . . . but the nebula remains . . . changeless in its very evanescence". The metaphor looks overworked even in the local stellar neighbourhood, where super-drive would be needed to dive through worm-holes of space in order to complete the itinerary.

For those willing to admit that suspension of disbelief necessary for enjoying science fiction, *Beyond the Moon* is a good read. It describes the Solar System, the nearest stars, and our Galaxy. Among the many commendable features is a surprisingly thorough survey of variable stars, a topic often skimped in popular writing. About one-third of the descriptive prose is a ramble through the extragalactic universe. Five appendices round off the journey, and there is also an outstanding index for those wishing to retrace their steps.

In no sense is *Beyond the Moon* a complete course in astronomy (neither author nor publisher claim that it is); rather it is like a series of extended essays in which the author is at his best when drawing out the themes that appeal to him particularly. His style and approach compares favourably with Asimov but lacks the incisiveness of Hoyle's writing at this same level.

Already *Beyond the Moon* has run through six editions in Italian, and it deserves to sell well in the English language. The translation and up-dating by D. J. K. O'Connell, S.J., of the Vatican, reads smoothly. Occasional lapses grate a little to anglophones (". . . the Protestant minister David Fabricius") and some analogy requires a good knowledge of Italian geography, such as knowing the distance by road from Rome to Genoa. This, however, is nit-picking, for O'Connell has done a fine job to this best-seller.

To whom is the book addressed? In the English language it must stand up to far more competitors than in its mother tongue. However, it seems admirably suited to that legendary and industrious hero of publishers' blurbs: the intelligent layman. It is a good account for the inquisitive but uninformed.

Simon Mitton

Simon Mitton is Secretary of the Institute of Astronomy, University of Cambridge, UK.

Structure of food webs

Food Webs and Niche Space. By J. E. Cohen. Pp.189. (Princeton University Press: Princeton, New Jersey, 1978.) Hardback £9.40; paperback £4.65.

ECOLOGISTS have been surprisingly myopic about food webs. They have described them; complained about how complicated they are (the "tangled knitting hypothesis"); noticed that the number of trophic levels is never very many; perhaps commented on the fact that some seem to be compartmentalised; and then got on with the serious business of studying single-species populations, interspecific competition, species-frequency distributions and what have you. Questions about the shape of entire food webs within which the ecological play actually takes place, have hardly ever been asked.

In this book Cohen assembles data on thirty-one food webs, accumulated over four decades of field work by other biologists. He claims no first-hand knowledge of the natural history of the plants and beasts involved, but seems to have been meticulously careful in reading what others have written about them. He then analyses the structure of these food webs, focusing on patterns of food sharing. Specifically, he asks whether the overlap in food utilisation among various kinds of organisms can be represented in one dimension, or more formally as an interval graph. Cohen's central conclusion is that community food webs are interval far more often than chance alone would provide.

A great deal of the book is taken up with deciding what chance will provide if she is allowed to design random food webs—to generate the hypothetical tangled knitting as it were. Thus, after setting the scene and gently introducing his reader to the jargon of graph theory and combinatorics in chapters one and two (it sounds more difficult than it is) Cohen moves to a description of his real food-webs in chapter three.

Chapter four contains a major statistical analysis of the properties of these webs, and reveals a number of hitherto unexpected properties over and above the surprising fact that most, perhaps all real webs are interval. For example, the number of kinds of prey is roughly three-quarters the number of kinds of predators in all the community webs; the number of links between the bottom of a food chain and the top is highly constrained but has different distributions in different webs; and as a multiple of the number of predators in the food web, the distribution of dietary overlaps is roughly uniform. There is much food for thought in these, and the other empirical patterns revealed in this part of chapter four.

The remainder of chapter four and part of chapter five discuss a series of model worlds; seven random webs constructed under a variety of assumptions and constraints. (A large part of chapter five is also devoted to the clinical diagnosis of a sick, or at best perverse random number generator, and says much about Cohen's meticulous care in the analysis of both data and models.) The main conclusion from this, the core of the book, is very clear. The high frequency of interval webs in the real world is extremely unlikely to have arisen by chance.

The remainder of the book is devoted (in chapter six) to asking why food webs should be interval. Cohen runs through three minor, and three major possibilities. I must confess that of the three interesting hypotheses, I found one so vague, and one so esoteric mathematically that I did not understand them. The third, focusing on the notion that there are dynamic constraints on the design of food webs, is to my mind the most interesting. Cohen's use of qualitative stability analysis merely scratches the surface, and indeed may be misleading as a large number of webs that are qualitatively stable turn out not to be locally stable when parameters are assigned and the models analysed properly. There is much scope for further work here. Finally, chapter seven flies away into some fairly esoteric air—the higher corners of Euclidean space and intersection graph theory—which may keep the pure theorists happy but is unlikely to lead to many empirically testable

hypotheses in the foreseeable future.

Obviously, the main empirical patterns which Cohen finds are only as good as the data on which they are based. Cohen himself is transparently honest about most of the weaknesses in his data, in the second part of chapter seven. There is no need to repeat them here. He also has some extremely perceptive remarks to make about data collection, and if the book does no more than encourage field biologists to collect new data on food webs that are specifically designed for and directed towards testing Cohen's hypotheses, it will have done a great service.

The main weakness in the data, however, is side-stepped by Cohen. In place of "species" he is often forced to use "kinds of organisms", and these may be very broad categories—for example, 'spiders" feeding on "collembola" in Figure 1. Without wishing to nit-pick, there are actually rather a large number of species of collembola, and even more sorts of spiders even in the one Canadian willow forest to which this particular web refers, and they are not usually trivial components of terrestrial ecosystems. Thus, properly constructed, this and I would guess most of the other food webs might look totally different. I do not know whether they would still be interval graphs, or how many of the other generalisations would still apply. This is decidedly not to say that Cohen should never have done the analysis in the first place. If ecologists always wait for perfect data nothing very interesting is going to get done. Nor will people ever suspect that there are theoretical problems worth solving with better data. But in this particular case we must definitely beware of the data, embrace the vision (Cohen claims that he got the idea in a dream), and take the book as a grand hypothesis. More than anything else, I hope that Cohen's contribution will stimulate others to look more seriously at the design of foodwebs.

John H. Lawton

John Lawton is a Senior Lecturer in Ecology in the Department of Biology at the University of York, UK.

Utilising phosphate resources

Phosphorus in the Environment: Its Chemistry and Biochemistry. Ciba Foundation Symposium chaired by R. J. P. Williams. Pp. 320. (Elsevier: Amsterdam, 1978.) $31.75; Dfl.72.60.

THIS symposium report contains 16 papers given by a diverse group of experts on the utilisation of the world's phosphate resources. The coverage is very wide; and chemical, biological, technical, ecological and economic factors are all discussed. Almost everything gets a mention: methods of birth control, water closets, marsh gas, nerve gas, toothpaste, detergents and even such mundane items as the current price of phosphate rock.

Although the pure chemistry content is somewhat limited, the book abounds with statistics, predictions and speculations about the future of mankind and how it will be linked to the availability of phosphorus. Reserves, consumption and the eventual fate of the element as utilised by man are critically discussed at some length.

Not unnaturally, there are differences of viewpoint and some contradictions in the data supplied by different authors. The symposium, nevertheless, makes a serious, valuable and timely contribution to its subject. Individual papers are of a high standard and the discussions are particularly informative.

An obvious message emerging from the symposium is that fertilisers will always constitute the most important use for the element, and if a rapidly expanding world population is to be fed, the conservation of phosphate supplies is already a matter requiring serious attention.

Phosphates are not in themselves harmful to life, and pollution problems arising from their use are local, often controllable, and are probably somewhat overrated.

With perhaps as much as three-quarters of the world's phosphate reserves lying in Arab countries, these reserves will undoubtedly, in the long term, prove to be their most valuable asset. Unlike oil, which can be replaced by alternative energy sources, phosphorus is irreplaceable in life systems.

It is pointed out that the complete natural cycle of phosphorus, unlike those of other major life elements, probably extends over many millions of years. This may be associated with a natural depletion of dry-land deposits and a temporary build-up of phosphate on the ocean floor. It is argued by some that the sudden and rapid exploitation of phosphorus compounds during the past 100 years, and the continued increase in their rate of consumption, will in a century or so lead to exhaustion of dry-land phosphate rock supplies.

It would therefore seem prudent not to worry unduly about detergents, but rather to concentrate on improving the present low efficiency of uptake of phosphorus from soils. This might be done, either by modification of existing fertilisers and soil systems, or by genetic engineering of suitable new plant species.

The eventual exhaustion of dry-land deposits, if this is to be believed, may one day necessitate direct sea-bed phosphate mining. Alternatively it may be possible to obtain food from suitable sea plants which could utilise phosphate directly from the ocean.

This is not a book for a newcomer, but it should appeal to all those with some previous knowledge of the subject who have wide interests in phosphorus science.

D. E. C. Corbridge

D. E. C. Corbridge is Lecturer in Chemistry in the University of Leeds, UK.

Microbial growth in extreme conditions

Microbial Life in Extreme Environments. Edited by D. J. Kushner. Pp. 465. (Academic: London and New York, 1978.) $40.50; £19.60.

THIS book needed to be written in view of the increasing significance of microbial growth in extreme conditions. The extreme environmental conditions chosen are principally physical ones: extremes of temperature, pressure, *p*H value, water and salt activities and radiation. Heavy metal toxicities are also included. This selection may be considered rather arbitrary, as it excludes the class of extreme environments caused by deficiencies of nutrients—for example, the trace elements, although excess of these elements is dealt with under toxicities.

A great interest of extreme environments is that the description of their effects may throw light on the effects of the normal or non-extreme environmental conditions. The special adaptations revealed may suggest possible new technological applications of the microbes. However, the main motivation of these studies has been to explain microbial ecology.

Many of the articles tend to be catalogues of facts which are useful for reference but not compelling reading. Also there is much turgid description of effects due partly to a lack of use of the quantitative parameters of growth such as specific growth rate and lag period.

The state of knowledge in the field seems to be in an early descriptive phase. The development of general explanatory principles will require much more study in depth of the problems from a whole organism viewpoint as well as molecular study. The reader will find in this work the growing points of the subject as well as a valuable reference source.

S. J. Pirt

S. J. Pirt is Head of the Microbiology Department at Queen Elizabeth College, University of London, UK.

obituary

James Craigie

PROFESSOR JAMES CRAIGIE, O.B.E., F.R.S., distinguished for his wide research in bacteriology, died in Edinburgh on 26 August 1978, aged 79.

He was born in 1899 and educated in Perth and at the University of St. Andrew's, where he qualified M.B., Ch.B. in 1923, and later Ph.D. and D.P.H. He was appointed assistant in bacteriology at the University in 1927. In 1931 he moved to the Connaught laboratories, University of Toronto, first as research associate, later as lecturer in epidemiology and, in 1946, Professor of Virus Infections. He returned to Britain in 1947 to join the staff of the Imperial Cancer Research Fund.

Craigie's introduction to microbiology was in the laboratories of St. Andrew's University under Professor W. J. Tulloch. Here he carried out work on the serology of typhoid and, with Tulloch, worked and published on the 'flocculation' reaction which occurred when vaccine lymph was mixed with vaccinia-immune serum. He continued these studies when he moved to Canada.

He showed that the 'flocculation' was due to a combination of agglutination of virus particles and precipitation of a soluble substance. Further studies with F. O. Wishart revealed that the precipitable substance consisted of two elements, one heat-labile (L), the other heat-stable (S). By ingenious techniques, separate antibodies against the L and S fractions were obtained; it became clear that a complex antigen (LS) containing both substances was present on the surface of the virus particle.

Following work over several years on the antigenic properties of typhoid bacilli, his most important contribution to knowledge was his work on typhoid bacteriophages. With K. F. Brandon he described in 1936 a phage specific for strains of *B. typhosus* (now called *Salmonella typhi*) containing what he described as Vi-agglutinogen. With C. H. Chen in 1938 he found other typhoid phages divisible into four types according to a range of properties. One of these (Type 2) developed a high selective lytic activity for the strain of *B. typhosus* on which it was propagated. Its activity could be regularly modified by changing the strain on which it was grown. With this research tool they found that typhoid bacilli could be divided into a number of types (24 were known by 1947). Strains which could be related on epidemiological grounds all fell into a single type. Here, then, was a method of great value in epidemiology, whereby evidence could be obtained as to whether individual infections could or could not be traced to a common source. The finding has had wide repercussions, for similar methods have been applied to the differentiation of strains of various salmonellas and other organisms.

During the Second World War Craigie was one of a number of workers concerned in developing a vaccine against epidemic typhus, and he also studied the serological relations of epidemic and murine typhus. He was a member of a joint United States-Canadian Commission which organised work on protection against rinderpest; it was feared that the enemy might try to introduce the disease into North America.

In 1946, W. E. Gye, the Director of the Imperial Cancer Research Fund, was touring North America to learn of developments in the cancer field, and Craigie accompanied him. Gye then persuaded him to return to Britain to the Fund's laboratories at Mill Hill with a view to becoming Director; this he did, succeeding Gye in 1949.

Gye, always a passionate believer in the virus theory of cancer held that the survival of mammalian tumour cells after freezing or drying indicated that cellular survival could be ruled out, and that a virus must be concerned. Craigie, however, showed that in frozen tissues, some particularly resistant cells might survive in a modified, 'paramorphic' state. He designed a very effective tissue mincer, and, more important, a cabinet for preserving transmissible tumours at −70 °C. As a result, numerous propagable tumours, especially those of mice, could be preserved for years; no longer was it necessary to passage them *in vivo* at frequent intervals, and much time and money were saved. His last work was on Tyzzer's disease of mice, a very troublesome infection in mouse colonies; he was able to cultivate the causative organism (*B. piliformis*) in fertile eggs.

Craigie was not really happy as Director, preferring to concentrate on his own research and in 1957 he was allowed to do this, Dr R. J. C. Harris assuming overall responsibility for directing the Mill Hill laboratory.

In 1929 Craigie married Margaret Fotheringham, who predeceased him. He is survived by two daughters. He was elected F.R.S., Canada, in 1946 and F.R.S., London, in the following year.

He will be especially remembered for the leading role he played in Canadian microbiological research, especially in the years before and during the war, when work on viruses was developing so rapidly. *C. H. Andrewes*

Harold Fletcher

DR HAROLD ROY FLETCHER, Queen's Botanist in Scotland, Regius Keeper of the Royal Botanic Garden, Edinburgh, from 1956 to 1970 died in Edinburgh on 27 August 1978. Born in Glossop, Derbyshire, he studied under Professors Weiss and Lang in Manchester and then moved to the Department of Botany of the University of Aberdeen as assistant to Professor Craib.

In Aberdeen his first responsibility—to his later amusement—was to teach plant physiology, but he soon turned to his main scientific interest, plant taxonomy and gained a PhD in 1933 for work on the Verbenaceae of Siam. At the same time he had a taste of plant collecting in the Faroes and Iceland and experience of the montane flora of Scotland for which he had an abiding affection.

In 1934 he moved to the Royal Botanic Garden, Edinburgh, then under the direction of Sir Willian Wright Smith and continued work on Asiatic floras gaining a DSc from Edinburgh University in 1939. Then followed a most rewarding period working on the rich collections of Chinese plants including a monograph of the genus *Primula* in collaboration with Sir William. At the same time he taught many university courses in botany, especially a course in taxonomy, which in its modern synthetic approach had no equal in Britain. Several strong modern currents in plant taxonomy in Britain flowered from these lectures.

During the years 1951–54 he turned his attention to horticulture—he had a rare eye for 'a good plant'—as Director the Royal Horticultural Society's Garden at Wisley in Surrey. And from this stemmed his activity in horticultural nomenclature as secretary for ten years of the International Commission for Horticultural Nomenclature

and Registration and his detailed enumeration of Rhododendron cultivars —all recognised by his award of the Royal Horticultural Society's Victoria Medal of Honour in 1956.

A year after his return to the Edinburgh Garden in 1955 he succeeded Wright Smith as Regius Keeper. Immediately his energy and enthusiasm were thrown into rejuvenating botany and horticulture in the Garden, in Edinburgh, in Scotland and further afield. He relentlessly badgered authority to give him the necessary resources and so built a new herbarium and library in 1964 and new glasshouses in 1967. The staff was enlarged and better equipped, the outlook in taxonomy and horticulture remoulded and the training of horticulturists recast.

His continuing interest in botany in Scotland found an outlet in stimulating the post-war revival of the Botanical Society of Edinburgh but his much wider interest in world botany led to his advocacy of Edinburgh as the site of the Tenth International Botanical Congress. For this Congress in 1964 he shouldered much of the planning as general secretary. This Congress placed botany in Edinburgh before a wide audience, but in addition by travels and lectures, especially in North America, he formed strong personal links worldwide—especially as President of the International Association of Botanical Gardens 1964–69.

His outstanding devotion to Botany in Scotland was recognised by his much cherished appointment as Queen's Botanist in Scotland in 1967, as an Honorary Professor of Botany in the University of Edinburgh in 1968 and by the award of honorary DSc's by Edinburgh in 1971 and St Andrews in 1972 and the Patrick Neill Prize by the Royal Society of Edinburgh 1971–73.

His interest in writing and horticultural and botanical history found an outlet in his *Story of the Royal Horticultural Society 1804–1968* in 1969 and (with W. H. Brown) *History of the Royal Botanic Garden, Edinburgh 1670–1970* published in time for the garden's tercentenary celebrations. Following his retirement in 1970 he wrote a detailed account of the plant collecting travels of Ludlow and Sherriff—*A Quest of Flowers.*

Throughout his professional life he had a deep love of the arts especially literature, music and painting of which he had a discriminating collection. In his retirement he found more time for these pursuits and served as Chairman and then as a member of the Governors of the Edinburgh College of Art.

Harold Fletcher was possessed of a most remarkable blend of science and the arts intermingled with a strong human compassion. His taxonomy was science tinged with art, his administration science guided by humanity. Throughout his life he drove himself hard and was loyally and very knowledgeably supported and encouraged by his wife Betty and by his son and daughter.

D. M. Henderson

John Marwick

Dr John Marwick, the leading palaeontologist in New Zealand during the middle decades of the twentieth century, died in Hawkes Bay, New Zealand, on 16 August 1978, at the age of 87.

The son of an Orkney wheelwright who migrated to Otago in the eighteen-eighties, Marwick was educated at Oamaru, attending Waitaki Boys' High School, where the rector, Dr R. H. Don, was changing the bias of the curriculum from classical to scientific, stimulating an interest in geology by classes and excursions such as were seldom available to secondary school boys. Entering the teaching profession, Marwick attended Otago University, graduating M.A. with first class honours in geology, and winning the Ulrich Memorial Medal in mineralogy and petrography, thereafter teaching in South Canterbury for two years before the outbreak of World War I. In the Middle East (1916–19) he served in the Camel Corps and N. Z. Mounted Field Ambulance, was twice wounded and was awarded the Military Medal.

On his return from the war, Marwick was appointed to work in palaeontology at the New Zealand Geological Survey (then under the Mines Department, but later a division of the Department of Scientific and Industrial Research). His research on Tertiary molluscs led in 1924 to the award of D.Sc. (University of New Zealand) for a revision of New Zealand Veneridae (Bivalvia).

Between 1921 and 1976 he published about 125 papers, reviews, and bulletins, his research continuing for 20 years after his official retirement from the Geological Survey (as Chief Palaeontologist) in 1952. He brought a sound objective judgement to the study of the systematics and biostratigraphy of New Zealand Bivalvia and Gastropoda. His revisions (illustrated by his own pen and ink drawings) of Struthiolariidae, Naticidae, Volutidae, Veneridae, Cardiidae and his monographs of Tertiary faunas, notably the Danian (Palaeocene) Wangaloan fauna, described jointly with the late Dr H. J. Finlay, earned John Marwick a world reputation as a student of fossil Mollusca.

With his colleague Finlay, who pioneered study of New Zealand Cretaceous and Cenozoic Foraminifera at the Geological Survey, Marwick was also responsible for defining and improving the standard sequence of local biostratigraphic stages that have been of great value in geological mapping in New Zealand. As a primarily Cenozoic palaeontologist, he regarded his occasional deviations into the Mesozoic as somewhat amateurish, yet his contributions include a standard reference work on New Zealand Triassic and Jurassic fossils. On the other hand he welcomed field work in a Geological Survey faced by the hard times and stringencies of the great economic depression, when Marwick was expected to assist in geological mapping and writing diverse reports with little relevance to palaeontology. He gained great satisfaction from descriptive geomorphology and stratigraphy, so that his publications in these subjects, for instance on the physiography and structure of the Rotorua–Taupo volcanic zone (with H.E. Fyfe) and on the geology of Te Kuiti district must always rank high among his achievements.

Dr Marwick might well have become Director of the Geological Survey, but although he served as Acting-Director on occasion, he preferred to remain in his palaeontology laboratory. On several occasions he lectured in geology at Victoria University of Wellington (then a college of the University of New Zealand) when his close friend Dr C. A. Cotton (later Sir Charles), well known for his geomorphology text books, was overseas on sabbatical leave, and he also served as external examiner for higher degrees. He took an active part in the Royal Society of New Zealand, which had awarded him the Hamilton Prize for beginners in science (1926) and elected him a Fellow in 1935; he was awarded both the Hector and the Hutton Memorial Medals of the Society for his contributions to palaeontology. He served on Council, was Honorary Editor of the Society's *Transactions* (1938–46) and was elected Vice-President (1950–52) but declined nomination as President.

Especially during the search for petroleum in New Zealand between the wars and after 1945, Marwick introduced many overseas geologists to New Zealand stratigraphy and he recalled with pleasure his excursions to fossil localities with Drs T. Wayland Vaughan, Axel A. Olsson, F. A. Singleton, and other visiting palaeontologists.

Jack Marwick was an unselfish man of high integrity, whose humanity and altruism made him a father figure to the many young geologists he helped in their careers; while his sincerity, good humour, and broad-based culture earned the friendship and respect of his colleagues.

C. A. Fleming